UNDERSTANDING THE ROLE OF EXPECTATIONS ON HUMAN RESPONSES TO AN AUTOMATED SYSTEM

A Thesis
Presented to
The Academic Faculty

by

Laura H. Barg-Walkow

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the School of Psychology

Georgia Institute of Technology
December 2013

UNDERSTANDING THE ROLE OF EXPECTATIONS ON HUMAN RESPONSES TO AN AUTOMATED SYSTEM

Approved by:

Dr. Wendy A. Rogers, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Richard Catrambone
School of Psychology
Georgia Institute of Technology

Dr. Francis T. Durso
School of Psychology
Georgia Institute of Technology

Date Approved: November 7th, 2013

ACKNOWLEDGEMENTS

I would like to thank Dr. Wendy A. Rogers, the Human Factors and Aging Laboratory (especially Sara McBride), and my family and friends for their support and guidance. This research was supported in part by a grant from the National Institutes of Health (National Institute on Aging) Grant P01 AG17211 under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Automation is defined as "machine execution of…functions that at one time could only be performed by humans" (Parasuraman, Sheridan, & Wickens, 2000, p. 286), and assists human operators with tasks to improve system performance. In general, automated systems make a wide range of tasks in a variety of domains safer and more efficient (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). However, automation is not always 100% reliable; it can commit errors by missing a signal in the environment or giving a false alert when no signal is present. Both types of automation errors influence human dependence on automated systems (Dixon & Wickens, 2006).

One factor that is known to influence dependence on automation is expected reliability of the automated system (Wickens & Xu, 2002). There are two common methods of introducing expected reliability of systems to users: explicitly telling operators what to expect or giving operators experience using the system (e.g., Chappell, 1997; Dzindolet, Pierce, Beck, & Dawe, 2002; Mayer, Fisk, & Rogers, 2008). However, there are gaps in the knowledge about human responses to automation, and how responses are differentially biased based on expectations. It is known that there are differences in human dependence on automation, but the differential impact (e.g., magnitude, duration of effects, nature of errors) of expectations is unknown. Specifically, the effects of explicit statements and initial exposures have not been directly compared over the course of time with consistent measures.

In this study, participants initially either received an explicit statement about or an exposure to the system reliability. For each introduction format, there were three levels of reliability: higher than (90%), lower than (60%), or the same as (75%) the actual system reliability for the rest of the experiment (75%), which isolated the effects of introductions. Following the introduction, all participants were tested using the same actual system reliability, dependent measures, number of trials, length of study, and error

types, frequencies, and timing. Holding these measures constant allowed for direct comparisons of influence of introductions across groups, including magnitude and robustness of effects.

Initially, there was an effect of expected level for explicit statement groups, whereas there was no effect of expected level for initial exposure groups. Over time, explicit statement groups had more stable perceptions of system reliability than the initial exposure groups. In general, perceived reliability did not converge to actual system reliability (75%) by the end of the study. Additionally, perceived reliability had a weak, but positive relationship with actual system use, whereas perceptions of system use (e.g., perceived dependence) had a strong, but negative relationship with actual system use.

Outside of initial effects seen with perceived reliability, there were few initial differences between expectation formats. Almost all groups tended to initially comply more than rely, with the exception of the initial exposure – lower-than group. Over time, level of expectation for initial exposure groups influenced reliance. There were no differences between expectation groups on compliance and dependence over time.

In general, dependence and compliance increased or stayed the same as time using the system increased. This pattern was also seen with reliance, with the exception of the initial exposure - higher-than group decreasing reliance over time.

Results from this study have implications for both theory and practice. The research findings both support and augment the existing conceptual model of automation. A better understanding of the differential effects of expectation format and introduced level of expectations can lead to introductions of automated systems that are best suited to the system's goals, ultimately improving system performance.

CHAPTER 1

INTRODUCTION

Technology is a ubiquitous component of many domains, ranging from the life-critical medical and military fields to use in homes. One important component of technology advances has been the adoption of automation, which is the "machine execution of […] functions that at one time could only be performed by humans" (Parasuraman, Sheridan, & Wickens 2000, p. 286). Automation usage can allow humans to focus on alternative tasks, which can lead to better overall system performance (Parasuraman et al., 2000). The shifting allocations of responsibilities can lead to safer, more efficient everyday tasks (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003).

However, automated systems are not perfectly reliable. Humans often serve as the final barrier before a system error—or series of errors—results in an accident. In these systems, human operators and automation operate as a "team", which includes emergent properties from the interactions (Bowers, Oser, Salas, & Cannon-Bowers, 1996). Therefore, it is important to understand all components of the system: automation, human operators, and the human-automation interaction.

The nature of the relationship between operators and automation is complex. McBride, Rogers, and Fisk (2013) identified different categories of variables, including: automation (e.g., system reliability); person (e.g., training); task (e.g., cost of verification); and emergent (e.g., workload). Research in human-automation interaction has been conducted in many of these categories; however, there is not yet sufficient research on the effect of expected reliability—and, particularly, its effect on perceived reliability—which is an emergent property of the interaction. A conceptual model of human-automation integration synthesized by Sanchez (2009) illustrates the relationships between prior knowledge of system reliability (here, expectations of reliability), actual automation reliability, perceived reliability, and automation use (see Figure 1). Investigating how expected reliability influences a human's overall use of the automation is important to understanding the human-automation relationship.

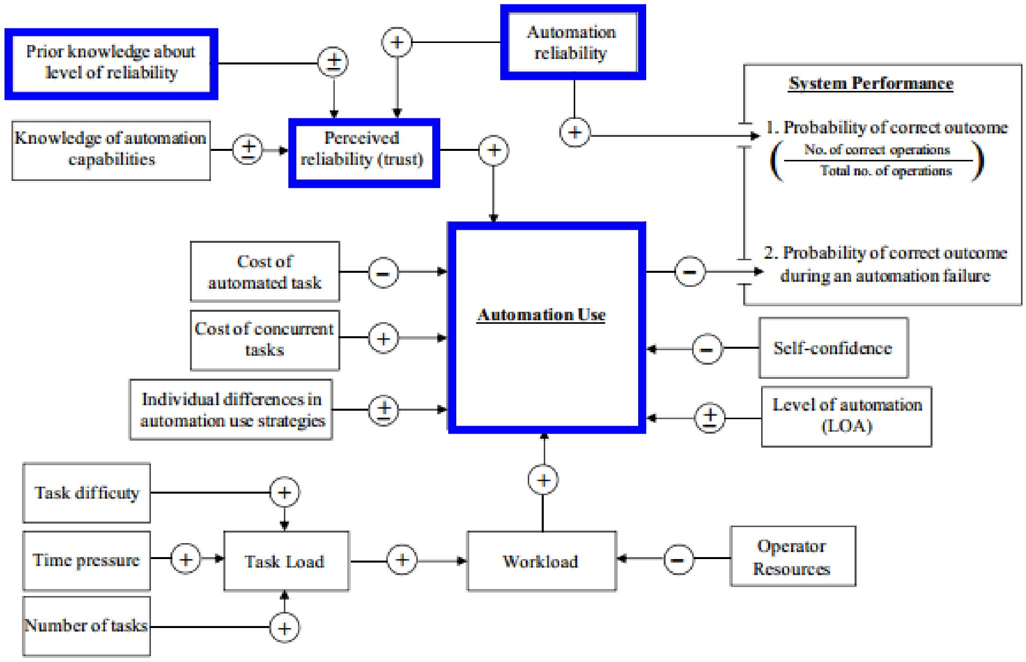

Figure 1. Conceptual model of human-automation integration (Sanchez, 2009). Variables important to the current study are highlighted.

Human Responses to Automation

Humans can have biases in their responses to automation when using a system, which could result in not using the automation optimally. Human usage of automation is generally discussed in terms of compliance, reliance, disuse, misuse, and dependence. Compliance and reliance occur when the human operator acts as the automation suggests, regardless of how correct the automation is. Dependence is the overall measure of matches between human actions and automation recommendations. Disuse and misuse, both types of errors, occur when there is a mismatch between human action and automation recommendations.

Compliance is “what the operator does when the automation diagnoses a signal in the world” (Dixon, Wickens, McCarley, 2006, p. 2). In other words, compliance is human action in response to an automation alert. For example, if a medical device is beeping to indicate that a patient’s vital stats are in a dangerous range, a healthcare professional can comply with the system by adjusting medications.

Alternatively, reliance is “what the operator does when the automation diagnoses noise in the world” (Dixon et al., 2006, p. 2). In other words, reliance is human inaction in response to silent automation. For example, if there are no alerts being issued through a medical device, a healthcare professional can rely on the system by not taking any action.

Dependence is the overall measure of matches between human actions and automation recommendations. Dependence can be thought of as the sum of human compliance and reliance with the system (Dixon & Wickens, 2004).

Disuse errors, commonly caused by false alarms, occur when a user does not comply or does not rely on the automation when the automation should be heeded (Parasuraman & Riley, 1997). For example, if a medical device is beeping to indicate that a patient’s vital stats are in a dangerous range, and the patient’s vital stats actually are in a dangerous range, it would be incorrect for a healthcare professional to not take action. This type of error can result in the neglect or underutilization of automation (Parasuraman & Riley, 1997).

On the contrary, misuse errors occur when a user incorrectly complies or relies on the automation when the automation directions should be rejected. For example, if a medical device is beeping to indicate that a patient’s vital stats are in a dangerous range, and the patient is actually stable, it would be incorrect for a healthcare professional to adjust the medications. This type of error can result in failures of monitoring or decision-making (Parasuraman & Riley, 1997). Factors that can contribute to misuse errors include: automation bias, automation reliability and consistency, obedience to “authority of automation", and workload (Mosier & Skitka, 1996; Parasuraman & Riley, 1997; Skitka, Mosier, & Burdick, 1999).

Automation Variables

Automation is not perfectly reliable, which impacts how humans should respond to automation (Table 1). A main way that the human-automation relationship is understood is through investigations of system errors (e.g., misses, false alarms). Patterns of automation errors can differ depending on the goals of the systems; for example, by changing the threshold for system alerts. A focus on detection of errors allows researchers to assess human responses to automation, understand why errors occur, and ultimately lead to the reduction of errors through future predictions of human behaviors (Kontogiannis & Malakis, 2009).

Automation Errors

When automation is not perfectly reliable, errors occur due to a mismatch between the state of the world and the alert state of the automation. Similar to classic signal detection theory, in the presence or absence of a signal, automation can be in the alert or not alert state. Alerts in the presence of a signal (hit) and no alerts in the absence of a signal (correct rejection) are both correct automation responses. Conversely, automation alerts in the absence of a signal (false alarm) and no alerts in the presence of a signal (miss) are both incorrect, as there is a mismatch between the state of the world and the automation response. Automation errors can influence human responses to automation; therefore, if a study is not directly comparing error types, it is important to hold timing and frequencies of the error types constant across participants.

Table 1

Human Responses to Automation

	Automation correct		Automation incorrect	
	Alert	**No Alert**	**Alert** (false alarm)	**No Alert** (miss)
Human – heeds automation	Comply (action)	Rely (no action)	Comply Misuse Commission	Rely Misuse Omission
Human – rejects automation	Non-comply Disuse Omission	Non-rely Disuse Commission	Non-comply (no action)	Non-rely (action)

Automation misses occur when a signal is present, yet the automation does not alert. Misses lead human operators to reduce their reliance on the system, which ultimately leads to more attention to the data and catches of rare events (Dixon & Wickens, 2004; Dixon, Wickens, & McCarley, 2006).

When automation gives an alert in the absence of a signal, false alarms lead to humans having both reduced compliance and reliance with the automation (Dixon & Wickens, 2004; Dixon, Wickens, & McCarley, 2006). False alarms were also found to have a detrimental impact on overall system performance (Maltz & Shinar, 2003). False alarms are typically more salient to operators than misses, and can lead to annoyance and distrust of systems, in addition to reduced reliance and compliance (Dixon & Wickens, 2004; Dixon, Wickens, McCarley, 2006).

Automation Reliability

The reliability of the automation affects human dependence on automation. Unsurprisingly, humans are more likely to depend on the automation when reliability is perfect or very high. As automation reliability decreases, other factors, such as perceived reliability, also decrease (Sanchez, 2009). In general, an automated system with reliability of less than approximately 70% (Wickens & Dixon, 2007) or less than 60-90% (Lee & See, 2004) is considered worse than not using automation.

Human Variables

Humans, similar to automation, are not perfectly reliable. Human biases, including mental models and exposures, shape expectations of and future responses to automation.

Cognitive Processes

A mental model is the human operator's conceptualization of a system's structure, including some understanding of how the system works and relationships between elements (e.g., Kieras & Bovair, 1984; Norman, 1983). These conceptualizations are generally thought of as dynamic—particularly changing with the availability of current information (e.g., Bibby & Payne, 1993; Eiriksdottir, 2011; Norman, 1987).

Mental models also serve as mechanisms for forming expectations and perceptions about systems. Two ways of being introduced to a system are knowledge by description, where a person obtains indirect, declarative information about the system, and knowledge by acquaintance, where a person acquires direct experience with the system (Russell, 1910). These types of introductions can also be thought of as described causal situations, where a person makes inferences from a linguistic description, and experienced causal situations, where a person witnesses the relationship between cause and effect when using a system (Shanks, 1991).

In studies by Bibby and Payne (1993, 1996), mental models formed through initial instructions persisted even after extensive use of the system. The studies suggest that mental models may be more impacted by initial introduction to the system than by dynamic updates to the system due to experience. However, other research has suggested that both judgments and biases in responses are similar following introductions by descriptions and experience, albeit through different associative learning mechanisms (Shanks, 1991; Wasserman, 1990).

An example of a mental model in the domain of human-automation interaction is the idea of “perfect automation”. In this schema, operators think of automation as more reliable than humans, but depend on it less following an error due to the proof of imperfection (Dzindolet, Pierce, Beck, & Dawe, 2002). Additionally, automation biases lead human operators to use heuristics: relying on automation in place of active information seeking and processing, leading to biases in human responses to automation (Mosier & Skitka, 1996).

Perceived Reliability

An operator’s perception of the reliability of the automation’s correctness is known as perceived reliability. This factor presumably mediates actual automation reliability and the human’s use of the system (e.g., Sanchez, 2009). In general, peoples’ judgment of perceived reliability for a system is lower than the actual automation reliability (e.g., Madhavan & Wiegmann, 2007; Sanchez, Fisk, & Rogers, 2004). Perceived reliability is positively correlated with misuse errors and negatively correlated with disuse errors (Dzindolet et al., 2002).

Perceived reliability is initially formed through humans’ expectations of an automated system. Expected reliability is the projected reliability of the automation’s actual reliability. Expectations have been found to bias attention and information selection (Bowers, Oser, Salas, & Cannon-Bowers, 1996; Stephan, 1985). People tend to seek out self-confirming information that aligns with expectations (e.g., Cantor & Mischel, 1977; Jamieson, Lydon, Stewart, & Zanna, 1987; Rosenthal, 1966; Rosenthal & Jacobson, 1968).

In the automation literature, expectations about system capabilities and functions influence dependence on automation (e.g., Cohen et al., 1998; Lee & Moray 1992, 1994; Lee & See, 2004; Muir, 1994, Riley, 1996; Wickens & Xu, 2002). Expected automation reliability has a dynamic relationship with human behavior of actual reliance on an automated system (Ezer, Fisk, & Rogers, 2007). Significant changes in perceived reliability over time occurred more frequently than associated changes in dependence on the system, suggesting that participants did not always change their behavior even if they believed the system' reliability had changed. Additionally, the first failure effect, wherein an operator shapes expectations—and calibrates perceived reliability—based upon the first failure of automation, is consistent with what would be expected from mental models of automated systems (Wickens & Xu, 2002).

There are two primary ways of altering human operators' expectations of system reliability. The first way that expectations can form is from outside information—similar to knowledge by description and described causal situations. This information can arise from sources including other users, supplemental materials, or the media, each of which is generally in the form of explicit statements. Expectations can also form through initial exposure to similar or identical systems—analogous to knowledge by acquaintance and experienced causal situations. Both formats for introducing automation can have an impact on human responses to automation (see Table 2).

Table 2

Studies Involving Automation Reliability Expectations for Explicit Statements (ES) and Initial Exposures (IE)

	Study	**Expected Reliability**	**Actual System Reliability**	**# Trials**	**Experiment Length**	**Responses Measured**	**Results**	**Frequency of Measurements**	**Effect Duration**
Explicit Statements	Bliss, Dunn, & Fuller (1995)	Initial: N/A Next: 75% vs. N/A	50%	Initial: 10 Next: 10	20 minutes	Response to alarms (compliance, implied dependence)	Higher expectation -> greater dependence; more compliance	Throughout study	N/A
	Dzindolet, Pierce, Beck, & Dawe (2002)	Positive framing (Pos) vs. Negative Framing (Neg) vs. No explanation	95%	200	Unknown	Intended dependence (binary yes/no)	Neg (then Pos, then no explanation) -> most intention to rely	Single measurement	N/A
	Madhavan & Wiegmann (2005)	"Novice system" (lower) vs. "Expert system" (higher)	70%	200	< 1 hour	Accuracy, Compliance, Dependence, Perceived reliability, Reliance, Sensitivity	High expectations -> initial over-dependence; later: under-compliance	Throughout study (except perceived reliability—single measurement)	Short (reliance); Longer (complia-nce for "expert")

N/A = Not Available

Table 2 continued

	Study	Expected Reliability	Actual System Reliability	# Trials	Experiment Length	Responses Measured	Results	Frequency of Measurements	Effect Duration
Explicit Statements	Madhavan & Wiegmann (2007)	"Novice system" (lower) vs. "Expert system" (higher)	70%	200	< 1 hour	Accuracy, Compliance, Dependence, Perceived reliability, Reliance, Sensitivity	High expectations -> conservative decision criteria; reduced compliance	Throughout study (except perceived reliability—single measurement)	Short (lower); Long (higher)
	Mayer, Fisk, & Rogers (2008)	Lower vs. Neutral vs. Higher	90%	160	80 minutes	Compliance, Error type, Reliance, Dependence	High expectations -> over-reliance and over-compliance	Throughout study	Short
	Mayer, Sanchez, Fisk, & Rogers (2006)	Lower vs. Neutral vs. Higher	90%	120	30 minutes	Error type, Dependence	High expectations -> over-depend; Low -> under-depend	Throughout study	Short (lower); Long (neutral, higher)

N/A = Not Available

Table 2 continued

	Study	Expected Reliability	Actual System Reliability	# Trials	Experiment Length	Responses Measured	Results	Frequency of Measurements	Effect Duration
Initial Exposures	Bahner, Hüper, &Manzey (2008)	100% vs. 80%	90%	Initial: 10 Next: 12	165 minutes	Dependence vs. Required dependence	High expectations -> depend less than low expectations	Throughout study	Long
	Chappell (1997)	100% vs. Neutral	90%	Day 1: 200 (100% only) Day 2: 200	Unknown	Signal detection (implied reliance)	Higher expectation -> initial over-reliance	Throughout study	Long
	Dzindolet, Pierce, Beck, & Dawe (2002)	90%	45% vs. 90%	Initial: 20 Next: 180	Unknown	Intended future dependence (binary yes/no)	System that remained reliable -> more intention to rely	Single measurement (end of study)	N/A

N/A = Not Available

Explicit Statements

As evidenced in Table 2, more studies have assessed the influence of explicit statements on perceived reliability and usage of automated systems. One study found that having explicit statements about higher system reliability led participants to greater dependence on the system (Bliss, Dunn, & Fuller, 1995). Initially, all participants experienced a system with 50% reliability. After practicing with this system, some participants were told that the system had a reliability of 75%, although the actual system reliability remained constant. In this case, being told a system has a reliability (75%) greater than that of the actual system reliability (50%) led participants to comply with the alarms at a higher rate than participants who had not been told the system's reliability.

In a study with participants using a luggage-screening task with a 70% reliable automated system, Madhavan and Wiegmann (2005) found that setting expectations high led to an initial over-reliance and over-compliance. Although this effect was gone by the end of the experiment in terms of over-reliance, participants in the high expectation group under-complied with the system over time. Additionally, setting expectations high led participants to set more conservative decision criteria with greater shifts towards optimal beta compared to participants with low expectations (Madhavan & Wiegmann, 2007). Perceived reliability was judged as lower than actual system reliability for both the 70% and 90% reliable systems. In both cases of system reliabilities, high expectation groups' ratings of perceived reliability were higher than the low expectations groups' ratings.

In a study with an 80% reliable system, inducing different expectations of system reliability also led participants to change their reliance patterns (Mayer, Fisk, & Rogers, 2008). The researchers found a short-lived effect where high expectations led to over-reliance and over-compliance.

Similarly, Mayer, Sanchez, Fisk, and Rogers (2006) found that high expectations persistently led to over-dependence whereas low expectations led to under-dependence.

The under-dependence in the low expectations condition was due to both low reliance and high compliance. Participants in the neutral group were most able to correctly adjust their dependence on the system by correcting both their compliance and reliance. Participants in the low expectations group had reduced dependence throughout the study. However, testing occurred over the span of 30 minutes; the longevity of these effects is ultimately unknown.

Additionally, in a study of automation aids, Dzindolet, Pierce, Beck, and Dawe (2002) found that when expectation was framed as negative (10 errors per 200 trials), intention to depend on the automation in subsequent trials was rated as higher than positive framing (half of participant errors = .5*[20 errors per 200 trials]) or when no explanation was given. This effect is perhaps due to the salience of non-perfect automation. In another study from the same paper, the authors again found evidence of framing statements impacting perceived reliability. However, data were only measured at one time point; the effect of framing over time is unknown.

In general, participants who were told that a system would have a higher reliability than the actual system reliability formed higher expectations of system reliability. These heightened reliabilities led to higher perceived reliabilities, resulting in participants over-relying on and over-complying with systems. However, this effect appears to be short-lived. But conclusions cannot be drawn with certainty because, as shown in Table 2, the studies using explicit statements used a variety of dependent variables and framing techniques over a range of exposure lengths.

<u>Initial Exposure to System</u>

As shown in Table 2, only a few studies have examined how exposure to a similar system influences perceived reliability as well as human usage of automated systems. One study examining the effect of initial exposures over time found that high expectations lead to over-reliance (Chappell, 1997). In the high expectations group, participants initially spent an additional day of testing using a 100% reliable system. On the second day of using the system, both groups used a system with actual reliability of 90%. Participants who had experience with the perfectly reliable system were more likely to over-rely on the system, making them less likely to detect the system failures due to their expectations. Additionally, participants with high expectations of system reliability exhibited a greater time lag before behavior adjustments than participants with low expectations.

Initial exposure to imperfect automation was also found to influence information sampling behavior (Bahner, Hüper, & Manzey, 2008). During training, one group of participants was exposed to automation errors, whereas the remainder of participants experienced only perfect automation. The experimental session included nine correct automation alerts followed by an error. No direct effect of training on error detection was found; however, the participants who were trained with imperfect information sampled significantly more information and were less dependent on the automation.

Additionally, a study by Dzindolet, Pierce, Beck, and Dawe (2002) altered actual system reliability partway through an experiment. Initially, all subjects experienced a system with reliability of 90%. After 20 trials, some of the participants were switched to a 45% reliable system for the remainder of the 200 trials, whereas the remainder continued experiencing the 90% reliable system. Participants then were asked whether they intended to depend on the automated system for their performance data. Participants with higher expectations of system reliability due to their initial exposure intended to depend more on the system in the future. However, these data were not presented in terms of dependence throughout the experiment; only a single measure of intention to rely on the automation in the future was measured at the end of the experiment.

In general, participants experiencing a system with a higher reliability than the actual system reliability formed initially high expectations of reliability. These heightened expected reliabilities led to higher perceived reliabilities, resulting in participants over-relying on systems. Participants who initially experienced less-reliable systems were able to better adjust their behavior to match the actual system reliability. In experiments with prior experience with the system, the effects of initial conditions of non-perfect expectations seem to be robust; however, more research is needed. As seen in Table 2, there were only three studies involving initial exposures, and there were some inconsistencies in measurements. The studies used a variety of dependent variables and framing techniques over a range of exposure lengths.

Explicit Statements and Initial Exposure to System

In both types of initial expectation formations—explicit statements and initial exposure to the system, high expectations lead to overreliance. Time also tended to change the patterns of dependence. Explicit statements, as opposed to initial exposures, had an effect on over-compliance. Additionally, the effects of expectations from explicit statements seemed to be shorter in duration than the effects of initial exposures to systems. The differences between studies in measurement frequencies and experiment durations make it challenging to assess differences in effect sizes and lengths with certainty.

To date, only one study (Dzindolet et al., 2002) has examined both formats. This particular study is limited in comparing the effect of introduction formats in that: the formats were not compared within the same experiment; the actual reliability of the system was not consistent; dependence was not measured over time; and reliance was measured only by intention to rely on the system in the future. There are no studies that directly compare explicit statements and initial exposure to the system. Directly comparing the initial expectation formations types and measuring human responses to automation over time will provide insights into the magnitude and duration of the effects of user expectancies.

Furthermore, studies examining the relative lengths and magnitudes of effects between these two types of expectation formations are limited. Only one study examined the effect of expectations over multiple days (Bahner, Hüper, & Manzey, 2008). This field of research is also limited in that the studies use different combinations of metrics to understand dependence, including reliance, compliance, signal detection, and perceived reliability. Not having consistent measurement variables reduces the ability to compare results and generalize findings.

Study Overview

The purpose of the research was to systematically assess how user expectancies—expectation format and level of expected system reliability—influence human responses (e.g., perceptions, compliance, reliance, dependence) on an automated system. Additionally, the research assessed both initial human responses and how these responses change over time. Finally, the study investigated the relationship between perceptions about the system and actual system use.

In the literature, there are gaps in the knowledge about the impact of different types of expectation formations over time, as well as how they will impact human responses to automation. As automation becomes ubiquitous, it is important to know how differences in introducing automated systems will affect human-automation interactions. Therefore, I investigated the effects of expectation formats and expected reliability levels both initially and over time on: 1) perceptions about the automation and 2) human responses to the automation (e.g., reliance, compliance, and overall dependence). Having a two-day study allows for exploring the effects of expectations on responses over time, including a break in system usage. Overall, the purpose of this study was to provide insights into how manipulations of expectation bias human responses to automation.

To answer these research questions, I systematically manipulated both expectation formation type and level of expectation using a simulated automated warehouse management system (AWMS). The AWMS operated in the context of dual-task scenario in which participants played the role of a warehouse manager responsible for 1) receiving packages coming into the warehouse, and 2) dispatching full trucks out of the warehouse. Automation is commonly used in the context of multi-task, dynamic operations. The system used in this study was specifically representative of the introduction of an automated system because no participants had prior experience with the specific task, or even similar tasks.

CHAPTER 2

METHOD

Participants

Participants consisted of 65 Georgia Institute of Technology undergraduates. The data from five participants were excluded; one participant experienced a computer malfunction, and one participant did not return for the second day of the study. Three more participants were excluded based on a combination of self-reported issues with testing (e.g., hand cramps, fatigue) and reporting taking medications known to interfere with concentration. Throughout the rest of this document, the data will be reported for 60 participants.

The 60 participants were between the ages of 18 and 23 (M = 19.80, SD = 0.21), 37 males and 23 females. They received three hours of credit for participation. All participants were required to be fluent English speakers to ensure participants could understand the instructions; 54 of the participants were native English speakers. The racial breakdown of participants included: 40 White Caucasians; 8 African Americans; 10 Asians; 1 Multi-racial and 1 Other/Unspecified.

Materials

Ability Tests

Participants' near vision and far vision were assessed using the Snellen visual acuity exam. All participants had at least 20/40 near and far vision, ensuring they were able to view the materials. The Digit Symbol Substitution (Wechsler, 1997) was administered as a measure of perceptual speed. The Reverse Digit Span test (Wechsler, 1997) was administered as a measure of memory span. Additionally, the Shipley

Vocabulary test (Shipley, 1986) was administered as a measure of verbal ability. These abilities tests were used to assess whether there were any underlying differences between groups.

Explicit Statement Descriptions

Expectancy was manipulated by providing participants with a written description of the automated warehouse management system with which they interacted (see Appendix A). The experimenter read the system description to the participant while the participant read along. The description framed participants' expectations such that participants either expected higher automation performance (90% system reliability), lower automation performance (60% system reliability), or the same-as automation performance (75% system reliability) as the actual system performance (75% system reliability). The system reliability was bolded and underlined in all passages. In addition, the description provided participants with information regarding system misses and false alarms. The explicit statement descriptions were adapted from the expectancy descriptions used by Mayer, Fisk, and Rogers (2008).

Questionnaires

Automation Experience Questionnaire

The automation experience questionnaire was administered to determine the level of experience participants have had with automation (see Appendix B; Johnson, 2004). The questions related to familiarity with automated devices and experiences with automated devices. Participants indicated which automated systems they have used, as well as aspects of ease or difficulty of use and trust in automation. Four of the seven questions were fill-in-the-blanks; two had instructions to "select all that apply"; one used a 5-point scale question to assess how much a participant trusted automated devices/systems, ranging from 1 (not at all) to 5 (completely).

Automation Attitudes Questionnaire

The automation attitudes questionnaire was administered to determine the attitudes and biases participants have towards automation (see Appendix C; Johnson, 2004). This Likert scale questionnaire consisted of 13 5-point responses, ranging from 1 (strongly disagree) to 5 (strongly agree). Participants compared how they felt about humans completing a task versus automation completing a task.

Automation-Induced Complacency Potential Questionnaire

The automation-induced complacency potential questionnaire was administered to determine individual differences in initial biases participants have towards automation usage (see Appendix D). This Likert scale questionnaire was originally developed by Singh, Molloy, and Parasuraman (1993), and was recently updated by Pop (2013) to reflect current technologies. The 20 items in this scale presented participants with scenarios involving different machines, for example "If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer than manual surgery." Each question was anchored on a 5-point Likert items, with responses ranging from 1 (strongly disagree) to 5 (strongly agree).

Demographic and Health Questionnaire

The demographic and health questionnaire, adapted from materials developed by the Center for Research and Education on Aging and Technology Enhancement (Czaja et al., 2006), was used to collect participant information (see Appendix E). The demographics portion of the questionnaire consisted of six questions to capture gender, age, race, and English-speaking background. Additionally, participants were asked to list all medications they were taking at the time of the study, because some medications may have an effect on performance.

Expectancy Questionnaire

The expectancy questionnaire evaluated participants' projections of automation performance in the automated warehouse management system (see Appendix F). The questionnaire consisted of three 5-point Likert-type items. The scale responses ranged from 1 (not at all likely) to 5 (extremely likely), as well as from 1 (automation works for the human) to 5 (human works for the automation). Additionally, there was a free response for the numeric value of perceived reliability of the automation. This measure was designed as an expectation check for the three levels of expectation (higher, lower, same-as) used in both expectation formats (explicit statement, initial exposure). Prior to beginning Block 1, all participants took expectancy questionnaires to provide a baseline of expectation.

Perceptions Questionnaire

At the beginning of each day's set of experiment blocks and after the completion of each block of trials, participants were asked to rate the perceived reliability of the automation, their trust in the automation, perceived compliance with the automation, and perceived reliance on the automation for the previous block (see Appendix G). This questionnaire consisted of three free response numeric answers of perceptions on a scale of 0-100% and one 5-point Likert-type item on trust. The response regarding how much a participant trusted the automation ranged from 1 (not at all) to 5 (completely). The questions were slightly altered for the initial perceptions questionnaire prior to beginning experimental Block 5; the questions asked participants to reflect on their perceptions about the system over the previous day, as opposed to the previous block. This questionnaire assessed how participants' perceptions of the systems and system usage change over time.

Simulated Automated System

The two types of expectation formations (explicit statements, initial exposure to system) were systematically manipulated in the context of a dual-task automated system. The automated system was a simulated automated warehouse management system (AWMS) developed by the Human Factors and Aging Laboratory and used in previous automation studies (Mayer, Fisk, & Rogers, 2008; McBride, Rogers, & Fisk, 2011). The AWMS operated in the context of dual-task scenario in which participants played the role of a warehouse manager responsible for 1) correctly receiving packages coming into the warehouse, and 2) dispatching full trucks out of the warehouse. The AWMS is representative of many automated systems in that the system is a dual-task, dynamic system with novel tasks. No participants had prior experience with the task.

Receiving Packages Task

In the receiving packages task, participants were presented with a barcode consisting of a string of four symbols (see Figure for a screen shot of the Receiving Packages task). Each barcode represented a package that had been delivered and needed to be inventoried. In addition to this individual barcode, the participants were presented with a list of barcodes, with a goal of matching the individual barcode to one of the barcodes on the list. Participants navigated through the list of barcodes using the up and down arrow keys, and used the key labeled "Receive" (number pad 0) to select the barcode. Participants had seven seconds to make each selection. After seven seconds, or after a participant selection—whichever occurred first—the next individual barcode and list of barcodes appeared.

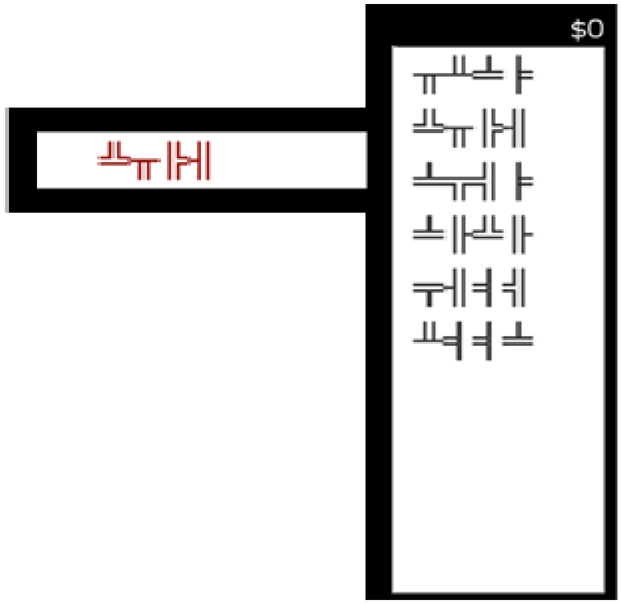

Figure 2. Screen shot of the Receiving Packages task.

Dispatching Task

Simultaneous to the receiving packages task, participants were completing the dispatching trucks task (see Figure 3). The automated aid alerted the participant when the truck had reached full capacity, although this was not perfectly accurate. The alert consisted of a red text box at the top left of the screen, viewable during all tasks without obscuring any portion of any task. Except for the Manipulation Block administered to participants in the initial exposure conditions, the system operated at 75% reliability. When participants were alerted to dispatch the truck, they could decide to dispatch the truck by pressing a key labeled “Dispatch Truck” (shift key), or ignore the automated aid. They could also verify the automation’s suggestion and view the interior of the truck by pressing a key labeled “View Truck” (spacebar). For verification of the automated aid, participants were required to press and hold down the “View Truck” key for at least two seconds, which blocked the receiving packages task from view and disabled all keys. There was also a two-second delay before the interior of the truck appeared, taking away from time on the receiving packages task. Participants were able to view the interior of the truck in this manner at any time, not only when the automation gave an alert.

Trucks loaded at a random rate, filling in between 12 and 22 seconds. Therefore, participants were not able to estimate when a truck should be fully loaded. All participants managed 20 truck loadings in each experimental block. When the truck was filled and the notification had been provided, participants had 10 seconds to dispatch the truck. If the truck was not dispatched within 10 seconds of being filled, the truck was considered overloaded and participants were penalized. If the truck was dispatched before it was filled, participants were also penalized.

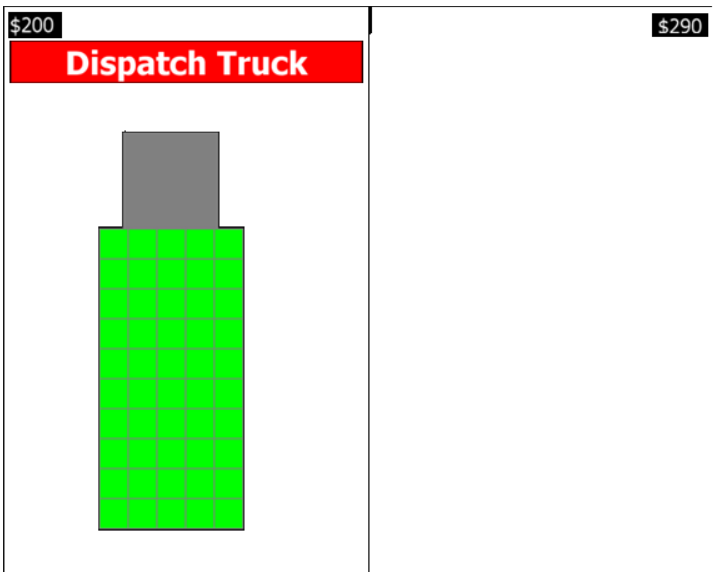

Figure 3. Screen shot of the Dispatching Trucks task. More specifically, this represents what is seen when viewing the shipping task (and not the receiving task) during an automation alert.

The automation was capable of making errors that were systematically programmed to be either a miss or a false alarm. A miss occurred when the automation did not recognize that the truck was full and, consequently, failed to send an alert. A false alarm occurred when the system notified the participant that the truck was full before it actually was, and, consequently, sent an alert too soon. Automation error types and frequencies, as well as the timing of errors within the blocks, were controlled in the automated system, as detailed in the Procedure section.

<u>Point Scheme</u>

Participants earned points for correctly receiving packages and for dispatching fully loaded trucks. Points earned for each task were constantly displayed to participants, and a grand total was for the combination of the receiving and dispatching task points was displayed to participants at the end of each block. The point scheme rewarded both speed and accuracy.

Receiving Task

Participants earned 15 points for each correct response. Participants lost 15 points for every incorrect response. If participant exceeded the seven-second time limit, it was considered an incorrect response and the participant therefore lost 15 points. The faster participants performed the receiving task, the more packages that could be received and the more points that could be earned.

Dispatching Task

Participants received 100 points for dispatching a full truck. If participants sent a truck that was not full, they lost 200 points. If participants overloaded a truck, they also lost 200 points.

Design

Expectation format (explicit statement vs. initial exposure) and level of expectation (lower vs. same-as vs. higher) served as between-participants variables in this 3x2 factorial design. Repeated measures of the participants' performance and perceptions served as a within-participants variable.

Independent Variables

Participants were randomly assigned to one expectation formation type (explicit statement, initial exposure) and one level of expectation (lower, same-as, higher), creating a factorial design of six between-participant groups. The level of expectation was manipulated in an explicit statement of system reliability for participants in the explicit statement conditions, and the level of expectation was manipulated through a Manipulation Block for participants in the initial exposure conditions (see Table 3).

Table 3

Reliability Levels throughout the Experiment.

<table>
<tr><td rowspan="4">Explicit Statement Condition</td><td>Level</td><td>Explicitly Stated System Reliability</td><td>System Reliability for Experiment Blocks 1-8</td></tr>
<tr><td>Lower</td><td>60%</td><td rowspan="9">75%</td></tr>
<tr><td>Same-As</td><td>75%</td></tr>
<tr><td>Higher</td><td>90%</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td rowspan="4">Initial Exposure Condition</td><td>Level</td><td>System Reliability for Manipulation Block</td></tr>
<tr><td>Lower</td><td>58.3%</td></tr>
<tr><td>Same-As</td><td>75.0%</td></tr>
<tr><td>Higher</td><td>91.7%</td></tr>
</table>

Dependent Variables

The automation task was divided into eight experiment blocks, with 20 trials per block. Dependent variables included quantitative, time-stamped measures of human responses to automation (e.g., reliance, compliance, dependence), as well as subjective responses to the questionnaires (e.g., perceived reliability of the system, perceived compliance, perceived reliance, perceived dependence).

Reliance, Compliance, and Dependence

Reliance, compliance, and dependence were measured as the percent of events during which participants did not press the spacebar key. That is, the instances when participants did not check the automation. There were 20 instances per trial when a participant could rely on the system and either 17 or 18 instances per trial when the participant could comply with the system (note: the discrepancy is due to automation miss errors, see Appendix J). Reliance was measured as the percentage of times during the reliance (non-alert) phase that participants did not view the truck to check on the accuracy of the automation. Compliance was measured as the percentage of times during the compliance (alert) phase that participants did not view the truck to check on the accuracy of the automation. Dependence was measured as the overall use of the system regardless of whether it occurred during the reliance or the compliance phase. Spacebar presses within a phase constituted non-reliance or non-compliance regardless of the number of spacebar presses. We also examined the number of times the spacebar was pressed and with what duration for any given component of non-dependence.

Procedure

Figure 4 presents the experimental procedure. Participants first provided informed consent. Next, they completed the Snellen visual acuity exam, followed by the Demographics and Health, Automation Experience, Automation Attitudes, and Automation Complacency questionnaires.

Participants were then given a general definition of the automated system with which they were to interact during the experiment. Specifically, the definition read, "An Automated Warehouse Management System is a system that scans the inside of truck trailer, calculates the amount of space available in the truck, loads packages onto the truck, determines if the truck is full, and when the truck is full, notifies the Supervising

Warehouse Manager to dispatch the truck." The general definition did not provide participants with any indication of whether the automation would perform well or poorly.

Next, participants completed seven practice blocks to familiarize them with the receiving and dispatching tasks (see Table 4). The purpose of practice was to ensure that all participants could perform the tasks after having the same exposure to the system by setting a criterion of performance ability that had to be reached. Practice blocks were designed to familiarize participants with the receiving packages and shipping tasks, both when the automation was correct and incorrect.

Table 4

Set-Up of Seven Practice Blocks with Receiving and Shipping Tasks.

	Receiving Task	Shipping Task	# Trucks	Automation Errors?
Practice Block 1	X			N/A
Practice Block 2	X			N/A
Practice Block 3		X	3	No—100% correct
Practice Block 4		X	3	3 Misses
Practice Block 5		X	3	3 False Alarms
Practice Block 6	X	X	3	No—100% correct
Practice Block 7	X	X	3	1 Miss, 1 False Alarm

The first and second practice blocks consisted of only the Receiving Packages task, first without the time limit of seven seconds enforced, and then with the time limit enforced in the next block. In each of the first two blocks participants were required to reach a performance criterion before proceeding to the next practice block. The performance criterion was five times the points awarded for a correct response of the receiving package task.

The third, fourth, and fifth blocks consisted of only the dispatching task. In the third block, participants were exposed to three trucks, for which the automation was always correct. Participants were informed that the system would be operating perfectly, which would not be the actual system reliability for the upcoming experiment. The fourth block contained three trucks, all of which included a false alarm by the system. The fifth block consisted of another three trucks, all of which included a miss committed by the system. In both the fourth and fifth blocks, participants were informed that the system would be operating imperfectly, which would be the actual system reliability for the upcoming experiment.

The sixth and seventh blocks included both the receiving and dispatching tasks. In the sixth block, the automation performed without error. Participants were informed that the system would be operating perfectly, which was not the actual system reliability. On this block, the participant was presented with three full trucks and did not have to reach a performance criterion to move on to the final practice block. The seventh and final block consisted of three trucks, two of which included an error by the automation (one false alarm, one miss). Participants were informed that the system would be operating imperfectly, and not at the actual system reliability for the trial blocks containing errors.

The experimental manipulation occurred in between the practice blocks and experimental Block 1. Participants in all levels of the explicit statement conditions read the explicit statement descriptions. To achieve the level of baseline reliability for the explicit statement group, written descriptions of the automated warehouse management system were manipulated to state the automation performed at a 60%, 75%, or 90% level of reliability (see Appendix D).

Participants in all levels of the initial exposure conditions experienced a Manipulation Block, for which they were responsible for dispatching 12 trucks. To achieve the level of baseline reliability for the initial exposure group, the automated warehouse management system performed at a 58.3%, 75%, or 91.7% level of reliability for the Manipulation Block. Half of the participants in each level of the initial exposure conditions experienced one more miss than false alarm (Group A), while the other half of participants experienced one more false alarm than miss from the system (Group B), to account for differences in error presentations (see Appendix H).

Prior to beginning Block 1, all participants were administered the expectancy questionnaire and perceptions questionnaire to assess baseline expectancies. Following the stated manipulations between the practice blocks and Block 1, all groups experienced the same experimental conditions for the entirety of the study.

During the experiment, participants were responsible for dispatching 20 trucks per block, and they completed a total of eight blocks (see Figure 4). To test the influence of expectation formats over time and after a time break, the eight blocks were completed over the course of two sequential days rather than in one single day, with four blocks occurring on each day. Following each block, participants completed the Perceptions Questionnaire. On the first day of testing, participants completed four experimental blocks.

The AWMS's reliability was consistent at 75% across blocks. To achieve this level of reliability, there had to be an odd number of errors, so in any given block, there was one more miss than false alarm, or vice versa. To compare results throughout a day and across days, Blocks 1, 4, 5, and 8—the blocks at the beginning and end of each day—had the same amount of misses and false alarms, and Blocks 2, 3, 6, and 7 had the same amount of misses and false alarms. Within each block, the misses and false alarms occurred in the same orders for all participants to eliminate response biases based on alarm types, frequencies, and order of presentation (see Appendix I).

On the second day of testing, participants first completed the remaining Abilities Tests, as well as the Expectancy and Perceptions Questionnaires. They were then given the same instructions regarding the automated system. Participants completed four experimental blocks, with the Perceptions Questionnaire following each block. Upon completion of all experiment blocks and questionnaires, participants were debriefed, compensated, and thanked for their participation.

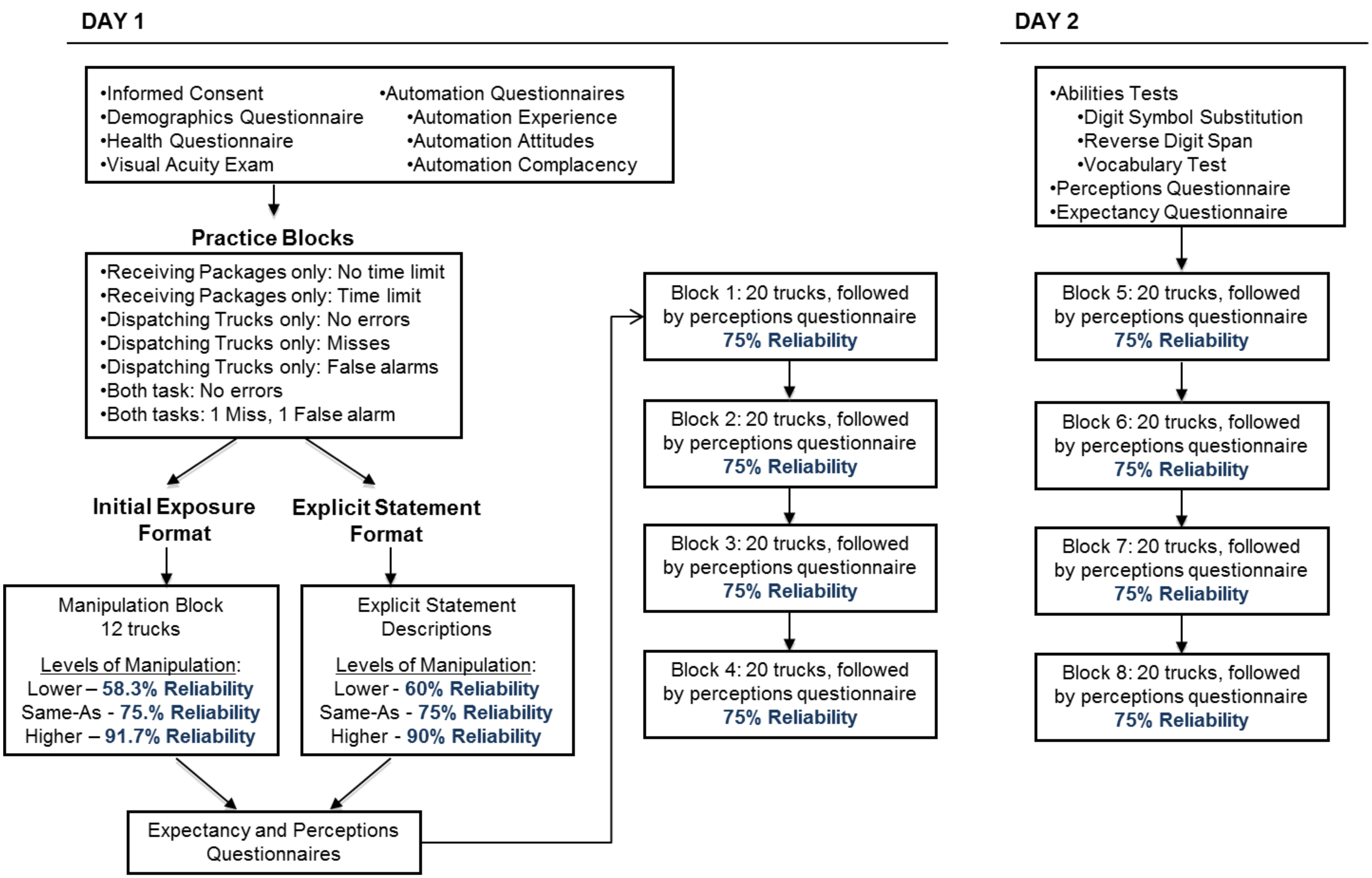

Figure 4. Experimental protocol

CHAPTER 3

RESULTS

Data Analysis

Unless otherwise noted, data were averaged across days instead of by experiment blocks to provide more stable estimates of perceptions and performance by group. The alpha level was set to .05 for all statistical tests. The data were analyzed using repeated measures analysis of variance (ANOVA), planned contrasts using t-tests, and Pearson correlations. Responses to individual questions on the automation attitudes and automation-induced complacency potential questionnaires are not presented; rather, total scores on each questionnaire were used in analysis. Data collected from the automation experience questionnaire are not presented.

Perceived Reliability

To assess how expectations initially affected perceived reliability both initially and over time, the data from the Expectations and Perceptions questionnaires were assessed at multiple time points (see Appendix K). These questionnaires were administered at the beginning of each day's set of experiment blocks (Baseline Day 1 and Baseline Day 2) and after each one of the 8 experiment blocks, giving 10 points of measurement. Initial data are from Baseline 1 only. Perceived reliability was reported on a scale of 0-100%.

System Use: Dependence, Compliance, and Reliance

The logged keystroke data of views of the system's automation both before and after automation alerts were assessed throughout each of the eight experiment blocks to assess how expectations affected actual usage of the system both initially and over time. Initial data is from Block 1 only. Dependence—and its subcomponents, compliance and reliance—were calculated on a scale of 0-100%.

Ability Tests

Table 5 depicts the means and standard deviations for the ability tests for each of the six groups: two expectation formats (Explicit Statement, Initial Exposure) x three levels of expected reliability (Lower, Same-As, Higher). Random assignment to groups did not yield differences of abilities. No significant differences were found on the Digit Symbol Substitution test, on the Reverse Digit Span test, or on the Shipley Vocabulary test (See Table 5). All participants scored within three standard deviations of the mean for each ability test.

Table 5

Ability Test Data by Group.

Group		Digit Symbol Substitution *(Max score = 100)*		Reverse Digit Span *(Max score = 14)*		Shipley Vocabulary *(Max score = 40)*	
Format	Level	Mean	SD	Mean	SD	Mean	SD
Explicit Statement	Lower	76.10	2.10	10.10	0.77	30.60	0.63
Explicit Statement	Same-As	77.00	3.23	9.90	0.72	31.30	1.18
Explicit Statement	Higher	67.10	3.82	8.80	0.61	31.20	0.90
Initial Exposure	Lower	67.10	2.22	9.20	0.48	31.30	1.38
Initial Exposure	Same-As	69.50	5.33	11.10	0.47	32.00	1.04
Initial Exposure	Higher	69.40	4.20	8.40	0.83	31.60	1.05
F-Value (5,59)		0.47		2.21		0.19	
P-Value		.80		.07		.96	

Initial Patterns of Perceptions and System Use

Initial Perceived Reliability

Following the manipulation of the levels of lower, same-as, and higher levels of reliability than the actual system reliability, it was expected that the initial perceptions of system reliability would map on to the expectations of the explicitly stated or initially experienced reliability (see Figure 5). One-sample t-tests were used to compare each group at the Day 1 Baseline to the manipulation level of expected reliability. The only group that was significantly different at baseline than the manipulated level was the initial exposure - lower-than group (M = 75.00, SD = 3.57), which was higher than the manipulated initial exposure (60%).

There was an effect of level ($F[2,54] = 4.78$, $p = .01$), with the higher-than level having a greater perceived reliability than the same-as level, which was greater than the lower-than level of expected reliability. There was also an interaction between format and level ($F[2,54] = 6.83$, $p < .01$), with the explicit statement groups having the same significant pattern of expectation level differences ($F[2,27] = 64.99$, $p < .01$) as the overall expectation level pattern, and the initial exposure groups having no effect of level. This means that different levels of explicit statements have more of an effect on initial perceived reliability than initial exposure groups. Additionally, despite having different introductions of expected reliability, there were no differences between initial exposure groups on perceived reliability, meaning that these groups were not as sensitive to changes in introduced levels of reliability when calibrating initial perceived reliability.

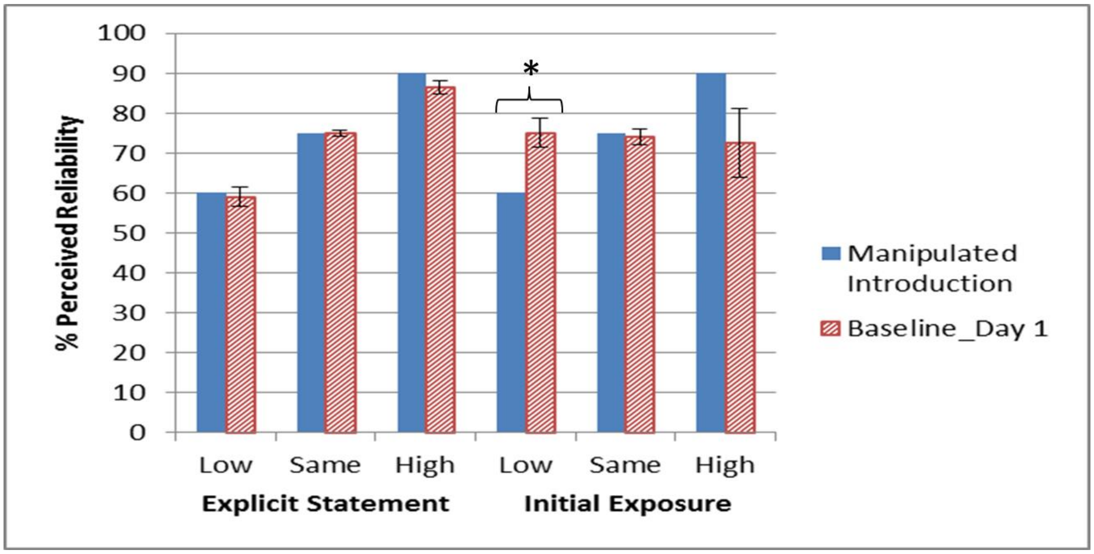

Figure 5. Perceived reliability at baseline day 1 vs. the manipulated expectations

Initial System Use: Compliance and Reliance

Expectation formats, levels of expectation, and their interaction did not have a significant effect on initial human responses of compliance and compliance (see Figure 6). We also analyzed differences between compliance and reliance for each group to test the findings in previous research that the expectation format of initial exposure leads participants to adapt their reliance more than their compliance on systems initially (Chappell, 1997; Dzindolet, Pierce, Beck, & Dawe, 2002). The initial exposure - lower-than group was the only group to have significant differences between initial compliance and initial reliance (t[9] = 2.64, p = .03).

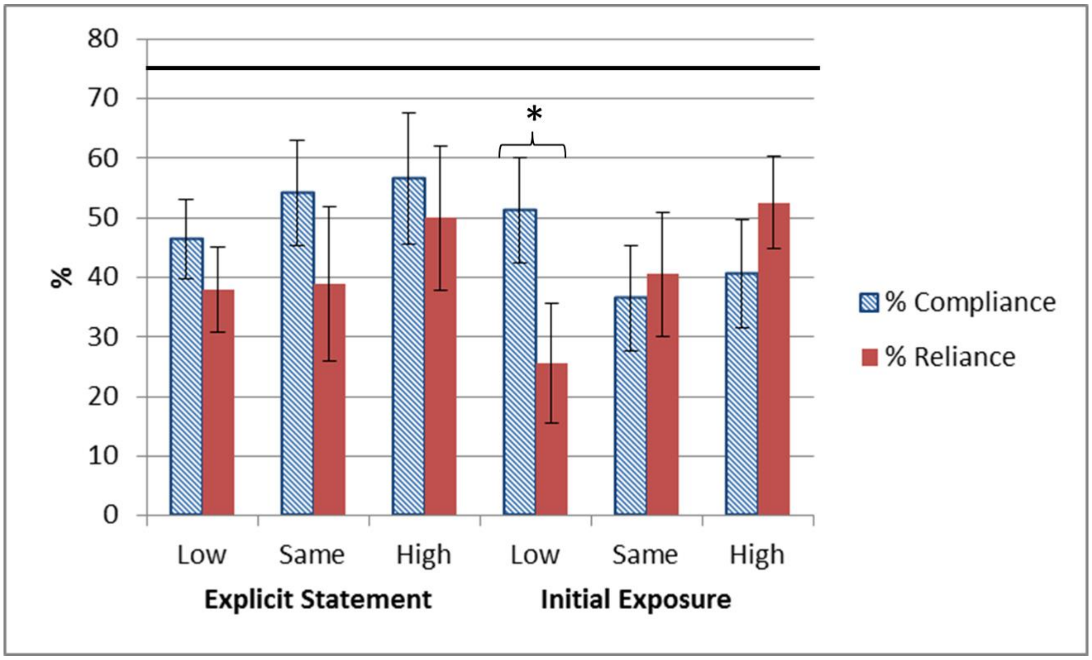

Figure 6. Initial reliance and compliance for each expectation format and level for Block 1 only.

Patterns of Perceptions and System Use over Time

When analyzing results from system usage, we uncovered substantial individual differences between participants. For example, the initial exposure – lower-than group ranged from scores of 0% to 100% on compliance over time (M = 52.23, SD = 30.06; see Figure 7). Due to this variability between participants, we decided to aggregate data across Day 1 (Blocks 1-4) and Day 2 (Blocks 5-8) for the remainder of the thesis—unless otherwise noted—to provide more stable estimates of perceptions and system usage.

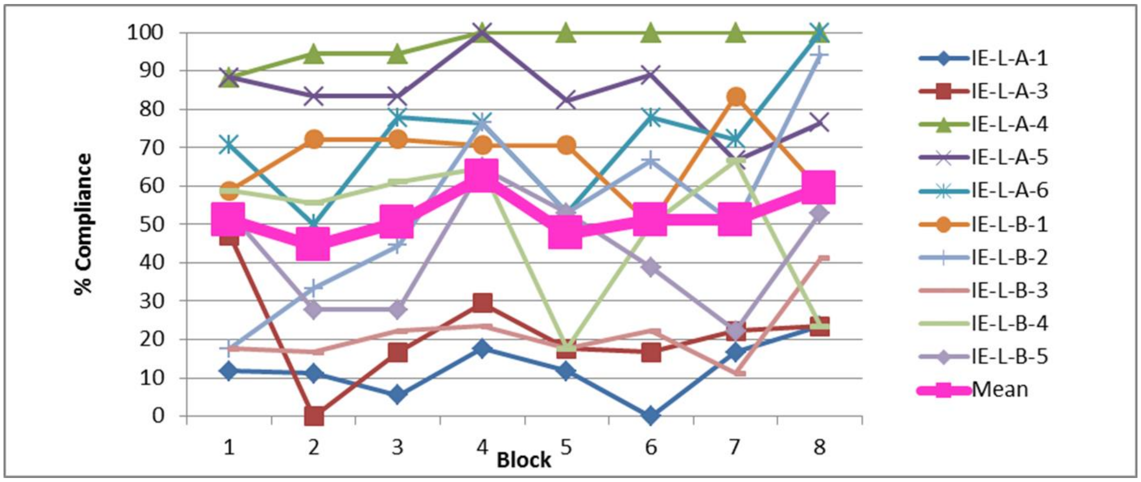

Figure 7. Individual differences on compliance for Initial Exposure - Lower-Than.

Perceptions of System Reliability over Time

To assess whether the effects of expectations on perceived reliability change over time, we compared the two baseline measures and conducted a day by format by level analysis. When comparing Baseline 1 to Baseline 2, we found that only the initial exposure-low group changed significantly ($t[9] = 2.61$, $p = .03$; see Figure 8).

There were no significant relationships for overall interactions between day, format, and level (see Figure 9). When split by format, level was significant explicit statement groups ($F[2,117] = 12.82$, $p < .01$; see Figure 10), with the higher-than group having a greater perceived reliability than the same-as group, which has a greater perceived reliability than the lower group. Level was also significant for the explicit statement groups ($F[2,117] = 4.22$, $p = .02$; see Figure 11), although in this case, the higher-than group has a lower perceived reliability than the same-as and lower-than groups. There was also a day by level interaction for the initial exposure groups ($F[2,117] = 3.22$, $p = .04$; see Figure 11), where the higher-than and lower-than groups increased from Day 1 to Day 2, while the same-as group decreased.

We also compared the later perceived reliability to the actual system reliability of 75% to assess how well the different groups were able to calibrate their perceptions of system reliability over time (see Figures 10-11). For this analysis, we only analyzed the final experiment block—as opposed to aggregating the data over the day—to better represent calibration at the end of the experiment. The explicit statement - lower-than group was the only group that was significantly lower than the actual system reliability, indicated by the star on Figure 10 ($t[9] = -3.87$, $p < .01$).

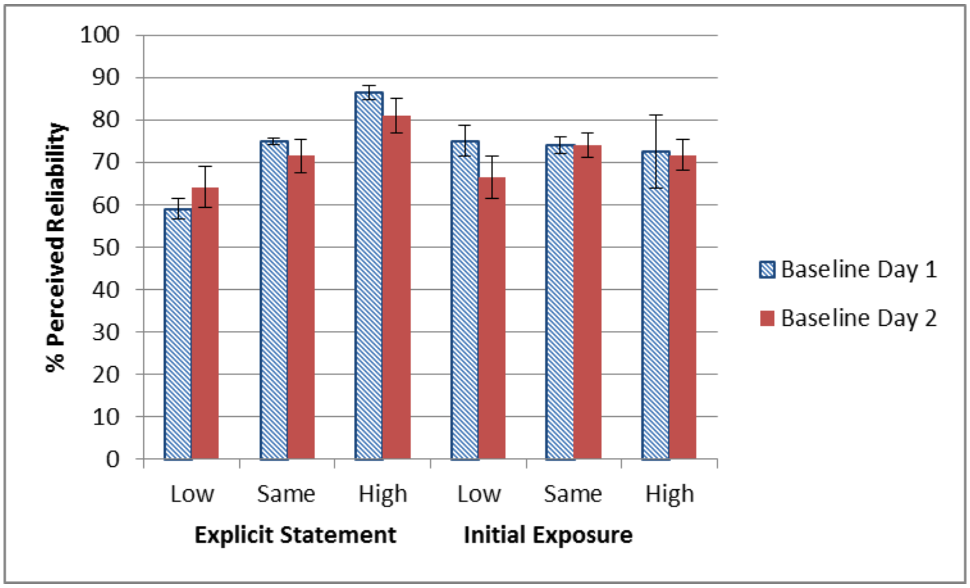

Figure 8. Baseline of perceived reliability at the beginning of each day.

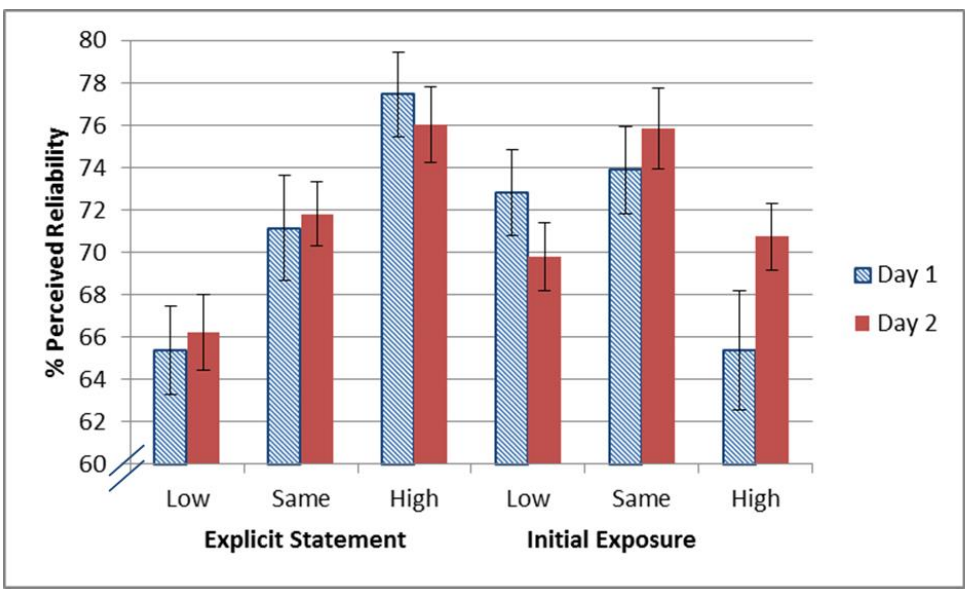

Figure 9. Perceived reliability each day.

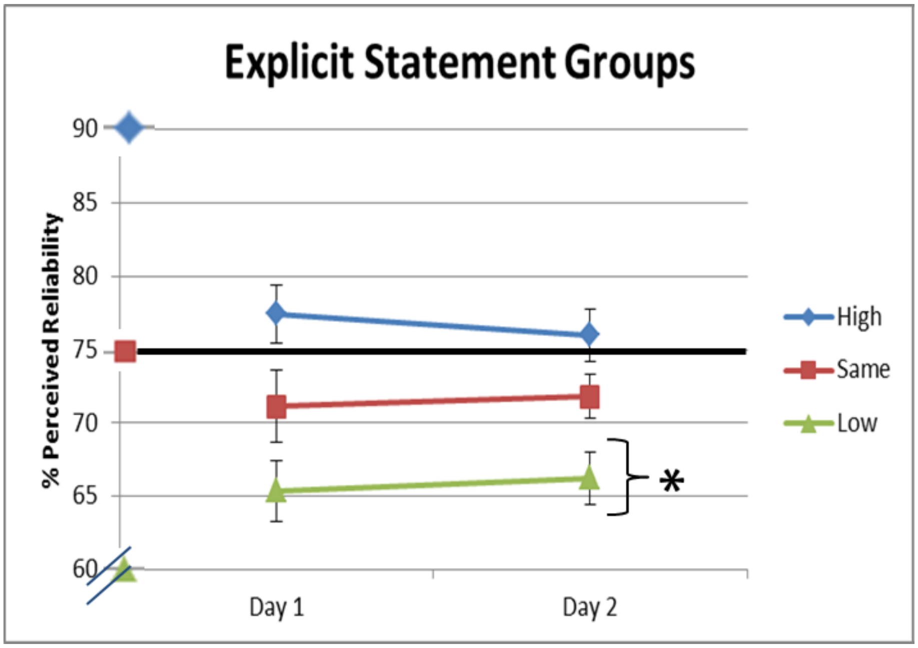

Figure 10. Perceived reliability by level and by day for Explicit Statement groups, including manipulated expected level of reliability (key shapes on y-axis) and actual system reliability (black line at 75%).

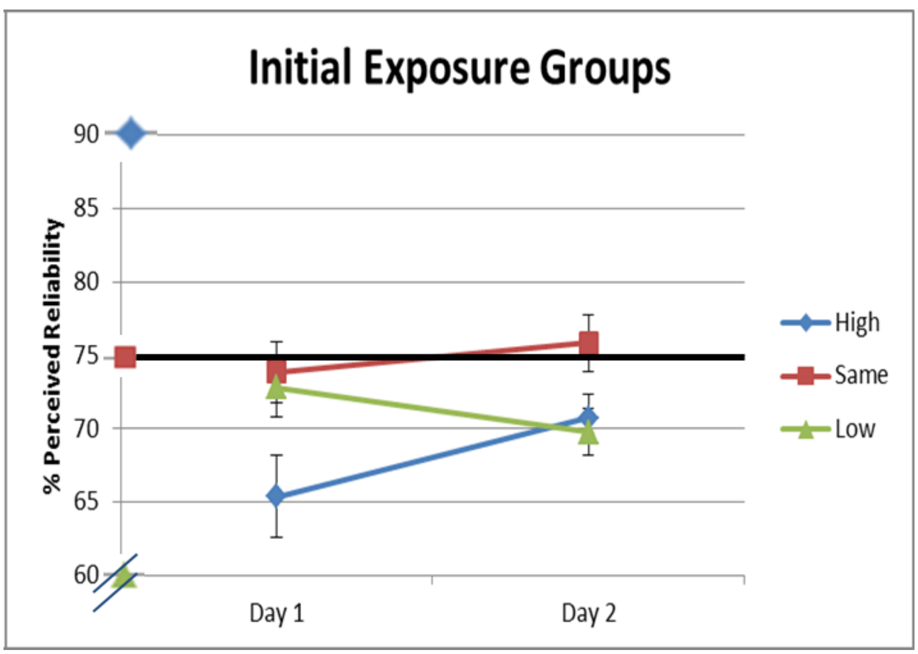

Figure 11. Perceived reliability by level and by day for Initial Exposure groups, including expected level of reliability (key shapes on y-axis) and actual system reliability (black line at 75%).

System Use over Time: Dependence, Compliance, and Reliance

There was not a significant correlation between dependence on the system and the automation attitudes questionnaire (*r*[58] = .10, *p* = .46). There also was not a significant correlation between dependence on the system and the automation-induced complacency potential questionnaire (*r*[58] = .09, *p* = .48).

To assess whether the effects of expectations on system usage change over time, we conducted a day by format by level analysis for dependence, compliance, and reliance (see Figures 12-14). For dependence, the interaction of days and levels was significant (*F*[2,234] = 3.51, *p* = .03; see Figure 12). Although all three levels increased from Day 1 to Day 2 for dependence, the trend was for the higher and lower groups to increase more across days than the same-as groups. Additionally, day was significant for dependence, with participants depending more on the system on Day 2 than on Day 1 (*F*[1,234] = 59.87, *p* < .01). There were no significant differences between groups on Day 1 only or on Day 2 only.

When split by format, the day was still significant for both explicit statement groups (*F*[1,117] = 24.77, *p* = .02), and for initial exposure groups (*F*[1, 117] = 35.71, *p* < .01). There were differences in the relationships of day, level, and format on system use when dependence was divided into compliance and reliance.

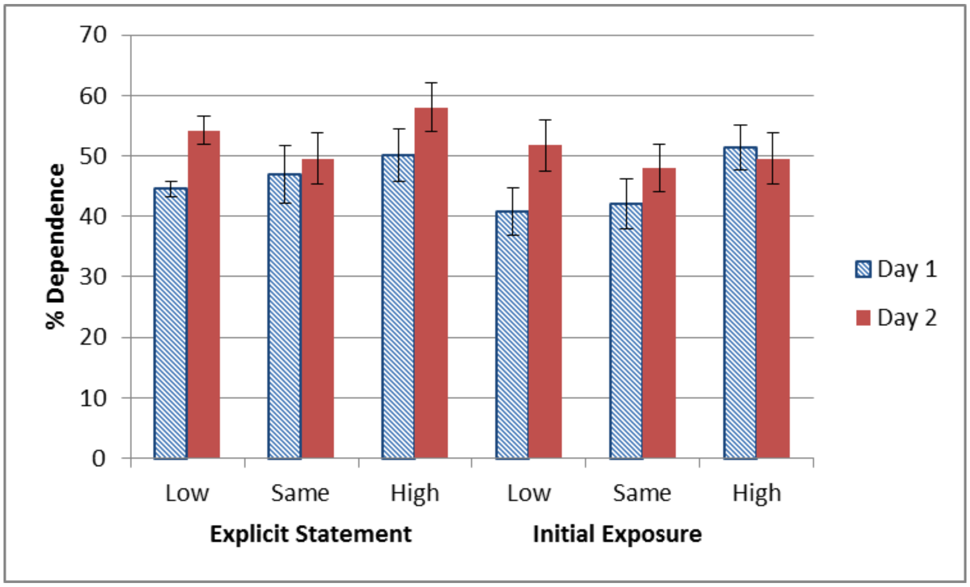

Figure 12. Dependence by day.

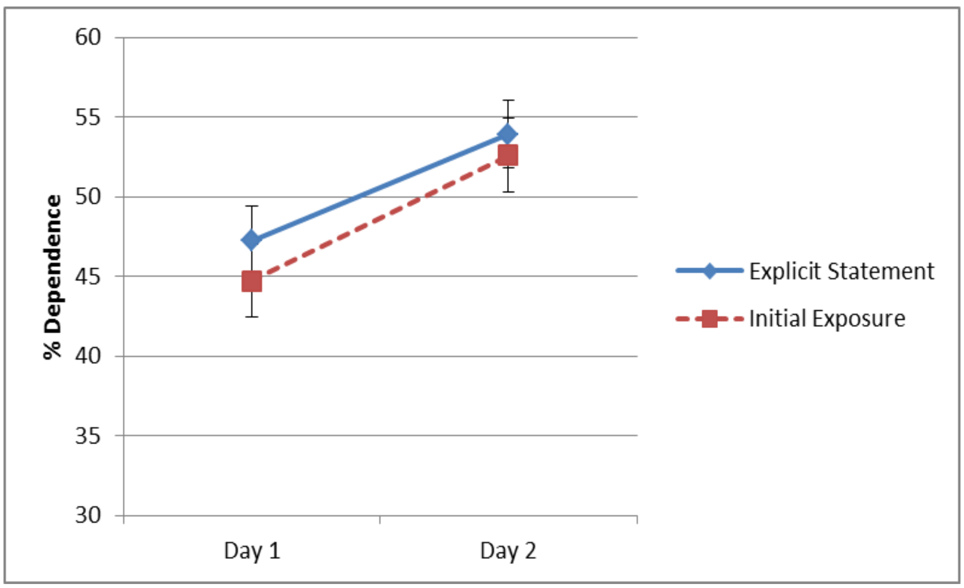

Figure 13. Dependence by format and by day, collapsed across level.

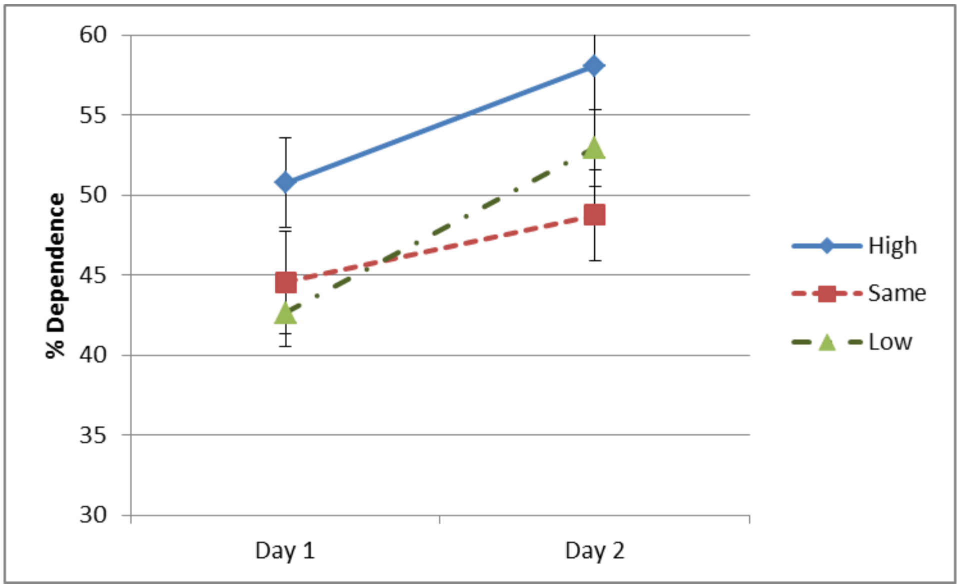

Figure 14. Dependence by level and by day, collapsed across format.

Compliance

On the first day of using the system, format was significant (*F*[1,234] = 5.41, *p* = .02; see Figures 15-17). On the second day of using the system, format was again significant (*F*[1,234] = 5.62, *p* = .02). For the effect of day, format, and level on compliance, the day was significant, with people tending to comply more on Day 2 than on Day 1 (*F*[1,234] = 12.36, *p* < .01).

When split by format, the day was still significant for both explicit statement groups (*F*[1,117] = 5.39, *p* = .02; see Figure 16), and for initial exposure groups (*F*[1, 117] = 7.34, p < .01; see Figure 17), with both groups complying more on Day 2 than on Day 1.

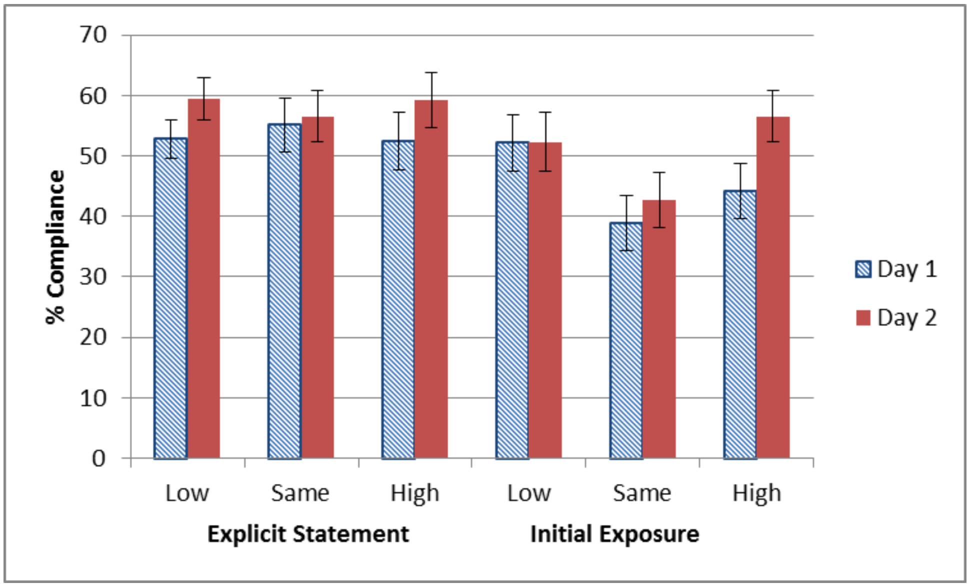

Figure 15. Compliance by day.

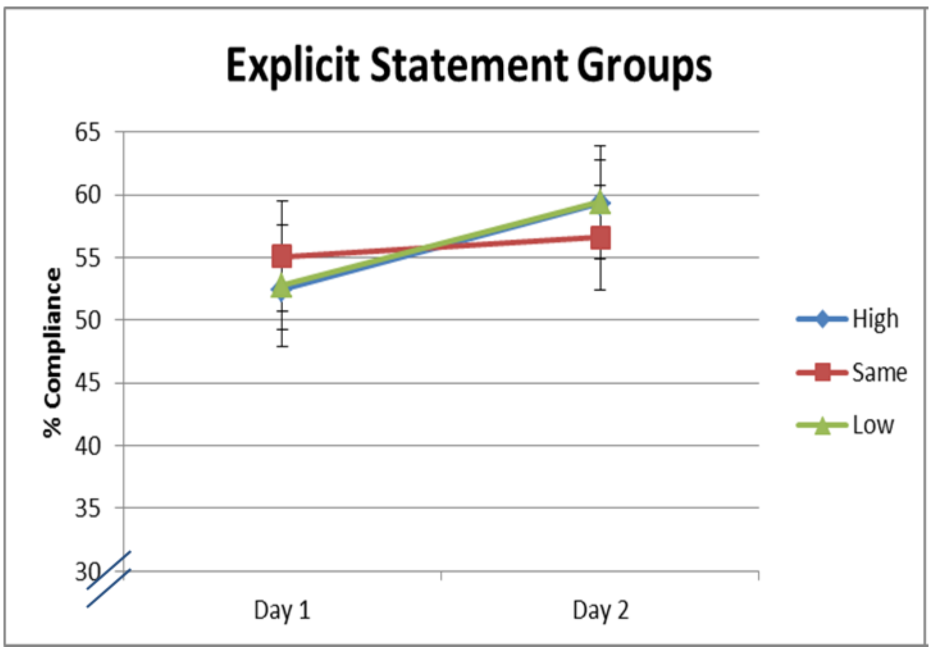

Figure 16. Compliance by format and by day for Explicit Statement groups.

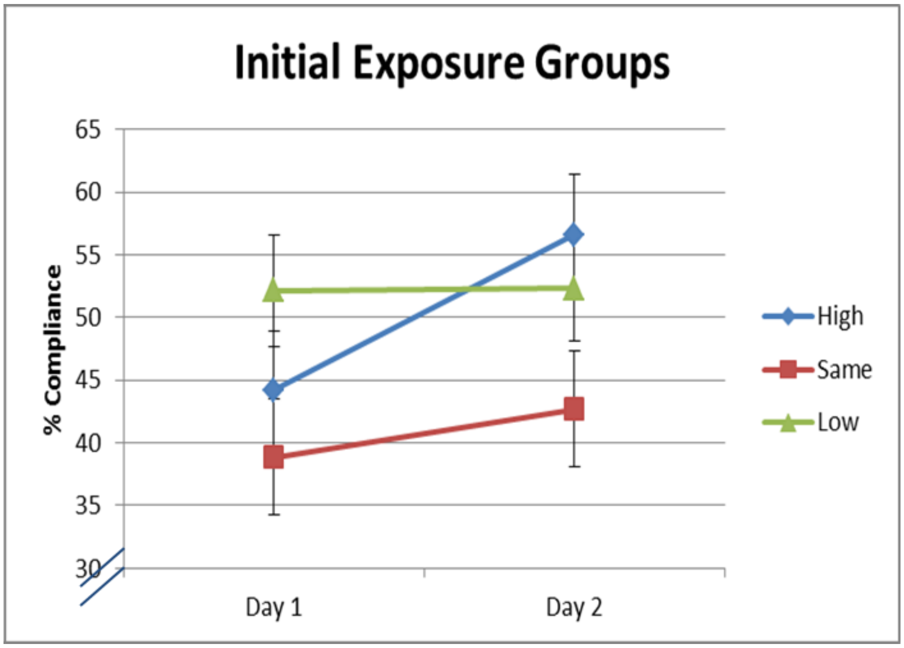

Figure 17. Compliance by format and by day for Initial Exposure groups.

<u>Reliance</u>

On the first day of using the system, level was significant (*F*[2,234] = 7.46, *p* < .01; see Figures 18-20). There were no significant differences on the second day of using the system. For the effect of day, format, and level on compliance, the day was significant, with people tending to rely more on Day 2 than on Day 1 (*F*[1,234] = 38.64, *p* < .01). Additionally, the interaction of day and level was significant, with the lower-than group increasing reliance more than the same-as and higher-than groups (*F*[2,234] = 5.46, *p* < .01).

When split by format, the day was still significant for both explicit statement groups (*F*[1,117] = 18.29, *p* < .01; see Figure 19), and for initial exposure groups (*F*[1, 117] = 20.56, *p* < .01; see Figure 20), with both groups tending to more on Day 2 than on Day 1. Additionally, for the initial exposure groups, level is significant, with the higher-than group relying more than the same-as group, and relying more than the lower-than group (*F*[2, 117] = 4.61, *p* = .01). There is also an interaction for initial exposure groups between level and day, with the higher-than group decreasing reliance from Day 1 to Day

2, whereas the same-as and lower-than groups increasing reliance from Day 1 to Day 2 (F[2, 117] = 4.79, p = .01).

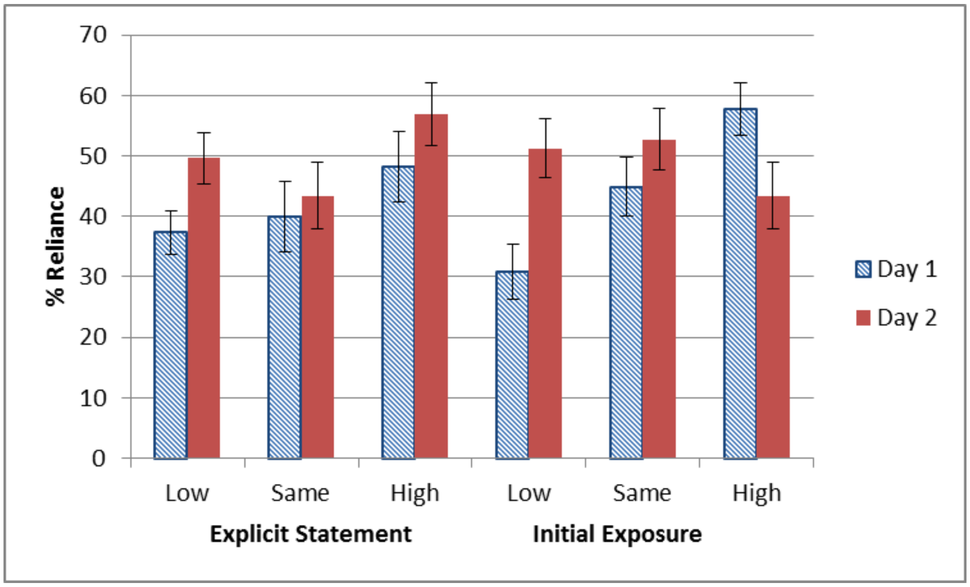

Figure 18. Reliance by day.

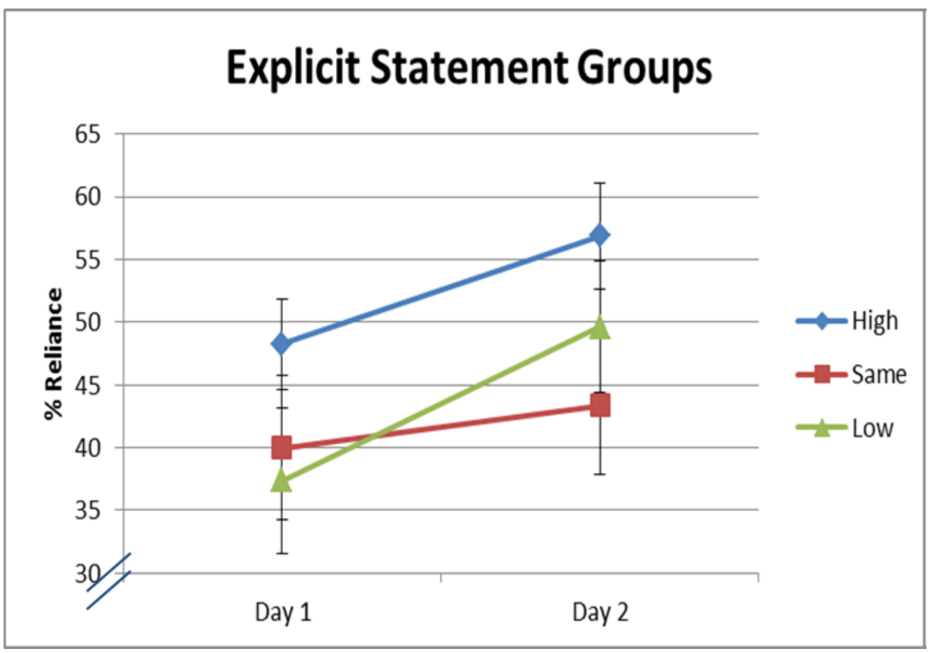

Figure 19. Reliance by format and by day for Explicit Statement groups.

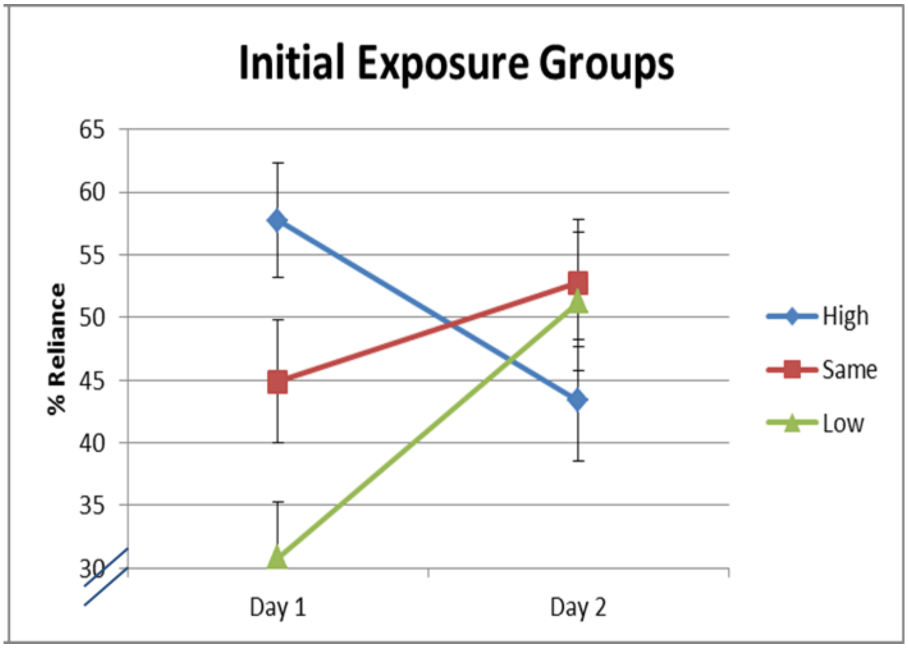

Figure 20. Reliance by format and by day for Initial Exposure groups.

Relationship between Perceptions and System Use

To assess the relationship between perceptions of the system and actual system use, Pearson correlations were calculated for each of the 60 participants over 8 experiment blocks, resulting in 480 points of comparison. First, we compared perceived reliability—which is used as predictor of automation use in Sanchez's (2009) conceptual model of automation—to actual system use (compliance, reliance, dependence). We found all of the relationships to be significant and positive (see Table 6). However, all of these relationships were weak according to Cohen's (1988) conventions to interpret effect sizes of Pearson correlation coefficients.

Table 6

Correlations between Perceived Reliability and System Use.

Variable 1	Variable 2	r(478)	*p*	Strength of Relationship
Perceived Reliability	Compliance	.14	< .01	Weak
	Reliance	.12	< .01	Weak
	Dependence	.16	< .01	Weak

We also compared actual system use (compliance, reliance, dependence) to direct representations of participants' perceptions of their system use (perceived compliance, perceived reliance, perceived dependence). We found all of the relationships to be significant and negative (see Table 7). Additionally, all of these relationships were strong according to Cohen's (1988) conventions to interpret effect sizes of Pearson correlation coefficients.

Table 7

Correlations between System Use and Perceptions of System Use.

Variable 1	Variable 2	r(478)	*p*	Strength of Relationship
Compliance	Perceived Compliance	- .75	< .01	Strong
Reliance	Perceived Reliance	- .50	< .01	Strong
Dependence	Perceived Dependence	- .69	< .01	Strong

System Performance

To answer whether expectations affected successful performance using the system, the logged data on points earned for each task with the system were assessed throughout each experiment block (see Figures 21-23). Participants scored between 915 and 4830 points in a single block on the receiving task (*M*=2790, *SD*=16.35), and between -1600 and 2000 points in a single block on the automated shipping task

(*M*=1449.38, *SD*=26.05). Total, participants scored between -595 and 6125 points in an experimental block (*M*=4158.38, *SD*=36.19).

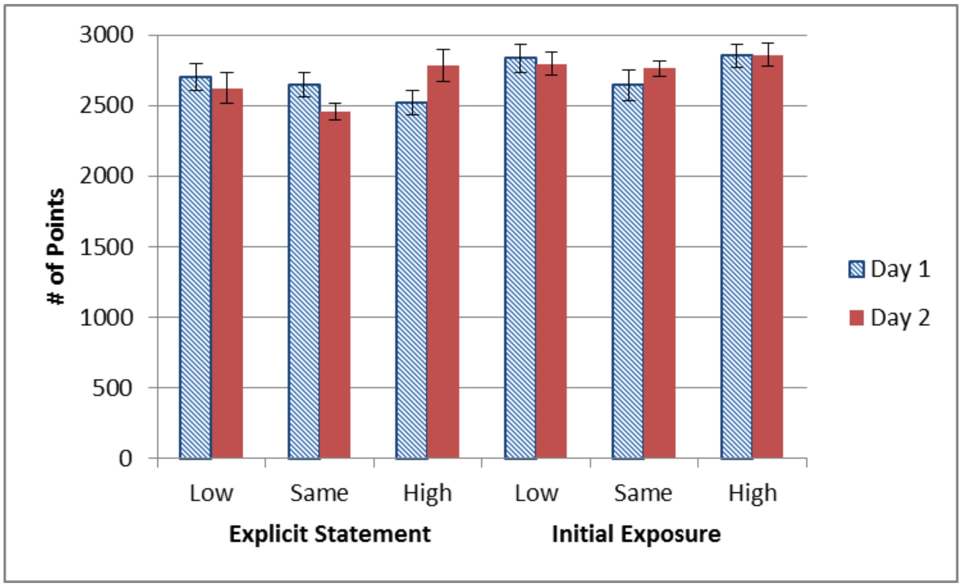

Figure 21. Points on the receiving task by day.

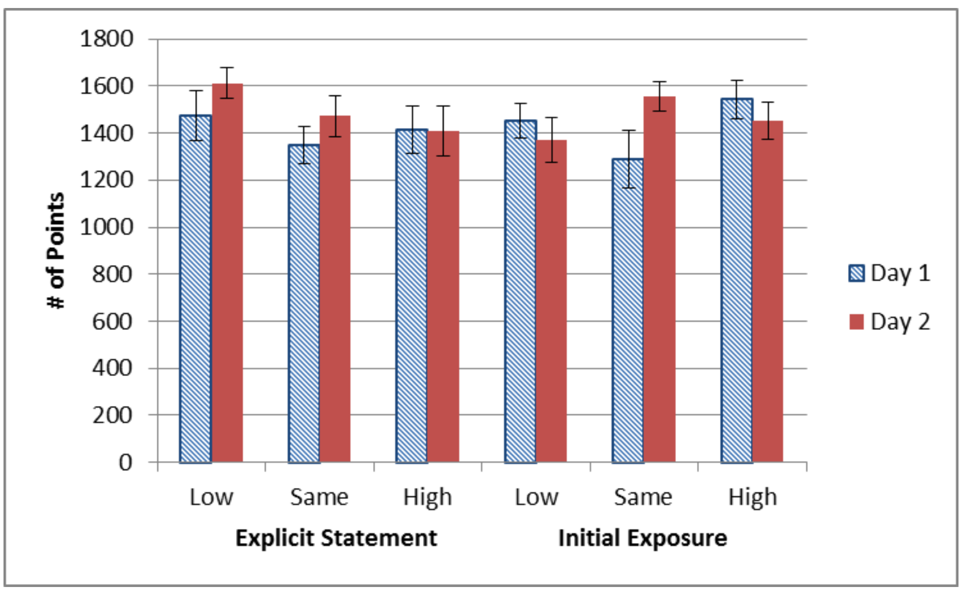

Figure 22. Points on the shipping task by day.

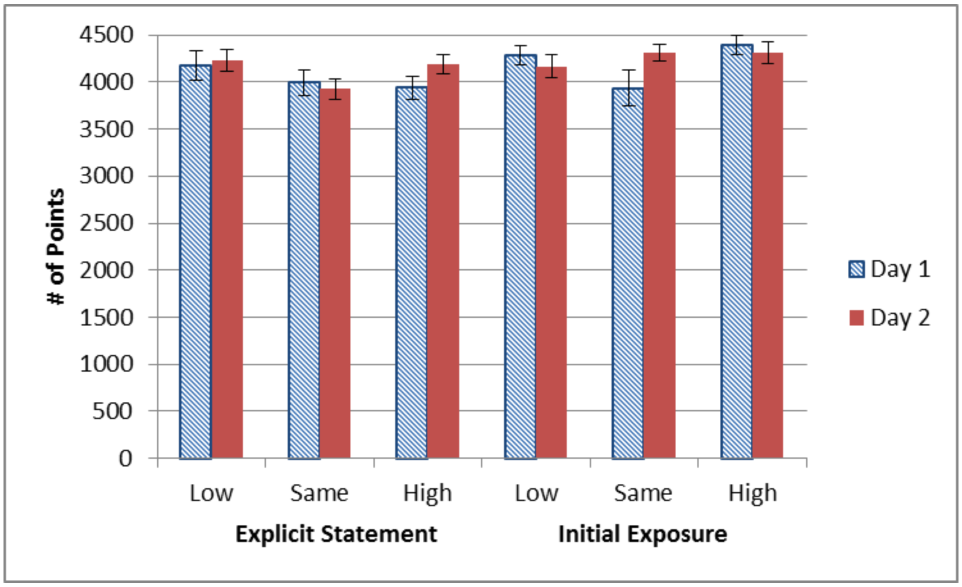

Figure 23. Total points earned by day.

Summary of Results

Initially, the expected level for explicit statement groups affected perceived reliability, whereas there was no effect of expected level for initial exposure groups on perceived reliability. Over time, explicit statement groups had more stable perceptions of system reliability than the initial exposure groups. Perceived reliability did not always converge to actual system reliability (75%) by the end of the study; half of the groups' perceptions remained significantly lower than 75%. Additionally, perceived reliability had a weak, positive relationship with actual system use, whereas perceptions of system use had a strong, negative relationship with actual system use.

Outside of initial effects seen with perceived reliability, there were few initial differences between expectation formats. Almost all groups tended to initially comply more than rely, with the exception of the initial exposure – lower-than group. Over time, initial exposure groups had a greater influence on reliance. There were no differences between expectation groups on compliance and dependence over time.

In general, dependence and compliance stayed the same or increased as time using the system increased. This pattern was also seen with reliance, with the exception of the initial exposure - higher-than group decreasing reliance over time.

CHAPTER 4

DISCUSSION

The purpose of the study was to determine whether user expectancies—expectation format and level of expected system reliability—influenced perceptions of and performance on an automated system both initially and over time. To that end, perceived reliability, dependence, reliance, compliance, and were investigated as a function of expectation formation type and level.

Perceived Reliability

Perceived reliability is presumed to predict automation use; initial perceived reliability can impact adoption of the automation, and perceived reliability can influence continued use of the automation over time. The literature suggests that the initial perceived reliability should match or be slightly lower than the introduced expectations (e.g., Madhavan & Wiegmann, 2007; Sanchez, Fisk, & Rogers, 2004). In the present results, initial perceived reliability was influenced by user expectancies for the explicit statement expectation format groups. However, there was no impact of different expectation levels for the initial exposure format groups. Potential reasons for not having an impact of expectation levels for this group include: 1) no sensitivity to changes in expectation levels with this introduction format; 2) too short of an initial exposure; 3) not trusting the manipulation; 4) practice effects; and 5) general bias towards assuming system reliability around 75%. Future work could focus on understanding why we did not find an impact of expectation level for the initial exposure format.

Over the course of the task experience provided in the study, level of expected reliability was related to perceptions of system reliability for both introduction format types. However, the patterns of these effects varied. For the explicit statement groups, the perceived reliability for the higher-than condition was greater than the same-as condition which was greater than the lower-than condition, mapping on to the introduced levels of reliability. For the explicit statement groups, this pattern remained the same throughout the experiment and did not change over time, suggesting that levels of expected reliability for the explicit statement groups had a robust impact on perceived reliability. For the initial exposure groups, the perceived reliability for the same-as condition was greater than the lower-than condition, which was greater than the higher-than condition, which did not map on to the introduced levels of reliability, although over time the perceived reliability converged for the lower-than and higher-than conditions by the second day. The perceived reliability for the higher-than condition for the initial exposure condition seems comparable to the first failure effect found by Wickens and Xu (2002), in that perceptions dropped following a decrease in automation performance from the introduction to the actual system usage. However, this effect has been shown with failures following perfectly reliable automation, whereas the introduction for this condition was at 90% reliability. Future research could investigate the range of introduced reliabilities for which the first failure effect holds true.

By the end of the experiment, we found that the perceived reliability for the same-as conditions for both formats and the higher-than condition for the explicit statement condition approached the actual system reliability, which is what we would expect, given that all groups experienced the same actual system reliability. For the explicit statement-higher group in particular, this means that the effect of introduction format was robust. It is unclear whether the perceptions for the same-as groups were due to the introduction, calibration to the actual system reliability, or to a combination of both factors. However, we also found that perceived reliability for the lower-than condition for both formats and the higher-than condition for the initial exposure condition remained lower than the actual system reliability by the end of the experiment. Having a lower expectation of system reliability had a durable effect on perceptions of reliability.

Additionally, data from this study suggest that demand characteristics of the experiment did not unduly influence participants' responses in perceptions of system reliability. Were participants answering in a manner consistent with experimenter expectations, we would expect that the perceived reliability for the different levels of expectation for the explicit statement groups to remain the same as the levels of introduction. However, by the end of the experiment, the higher-than group for the explicit statement introduction format was not different from 75%, whereas the group would be expected to be at 90% if they were responding based on demand characteristics.

Compliance and Reliance

Compliance and reliance are important to understand because there can be different consequences for heeding automation when it is incorrect and rejecting automation. In this study, we held false alarms and misses constant to ensure this was not driving changes in compliance and reliance rates. Based on the literature for introductions through explicit statements, it was expected for the explicit statement groups that introducing higher expectations would lead to higher dependence, whereas lower expectations would lead to lower dependence (e.g., Mayer, Fisk, & Rogers, 2008). However, we did not find differences between the levels of expected reliability for explicit statement groups on compliance or reliance, either initially or over time.

Studies on initial exposure introductions, such as Chappell (1997), suggest that higher levels of expectation lead to higher reliance, but not higher compliance. This could be due to the higher salience of false alarms over misses in the introduction. In our study, we did not find differences between the levels of expected reliability for initial exposure groups on either compliance or reliance initially, although the initial exposure group complied significantly more than they relied. Over time, there were no differences between initial exposure groups on compliance. However, for reliance, level of expectation for the initial exposure groups played a role; on Day 1, the higher-than group was greater than the same-as condition which was greater than the lower-than condition, although these group differences disappeared on Day 2. Overall, the explicit statement groups complied more with the automation than the initial exposure groups, which again is in accordance with previous work (e.g., Chappell, 1997).

Role of Individual Differences

We found very few significant differences between groups when analyzing results at the time granularity of experiment block. The lack of significant results could be for many reasons, including 1) no actual differences between groups, 2) variability, 3) low power, 4) restriction of range, and 5) sensitivity and reliability of measures. It is of course possible that there are no actual differences between groups. Additionally, the study included a small sample size of n=60, with 10 participants in each group. The measurement of variables in this study go from 0% (never) - 100% (always) for each of the perception and actual system use, which does not support restriction of range, although some participants did use the system 0% of the time whereas others used the system 100% of the time. There were 140 opportunities to comply, 160 opportunities to rely, and 300 opportunities to depend on the system, each of which were objectively measured. Variability, however, appears to be playing a large role.

Individual differences in the data were large and fairly stable over time, leading to high error for each measure. Similar patterns were also found in the data from other studies using this system (Mayer, 2006; McBride, 2010). Information about individual differences is generally not reported in the automation literature, although it could be a factor in other studies.

Implications for Theory

Results from this study inform Sanchez's (2009) conceptual model of automation (see Figure 24). In accordance with the conceptual model of automation, we found relationships of prior knowledge about level of reliability (Explicit Statements) and automation reliability (Initial Exposures) on perceived reliability. We also found a relationship of perceived reliability on automation use.

In addition to these findings, we found that automation use also influenced perceptions over time; this relationship was bidirectional. We also explored the relationship of other perceptions on automation use; we found that perceptions of automation use have a stronger relationship with automation use than perceived reliability. Our findings also suggest that perceptions may not mediate all of the effects of introductions (i.e. Explicit Statements, Initial Exposures) on automation use.

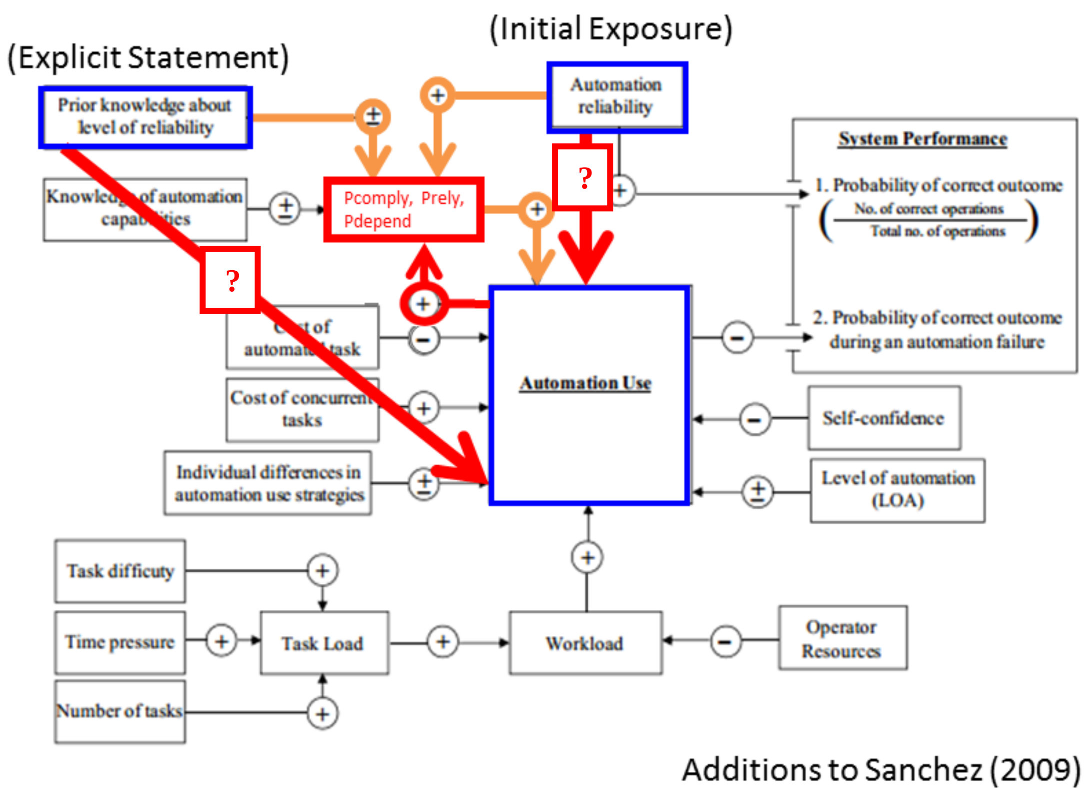

Figure 24. Reinforcement (shown in orange) and additions (shown in red) to the Conceptual Model of Automation.

Results from this study inform the relationship between self-perceptions and system usage. Generally, studies that measure perceptions focus on perceived reliability (e.g., Madhavan & Wiegmann, 2005; Madhavan & Wiegmann, 2005; Mayer, Fisk, &

Rogers, 2008; Mayer, Sanchez, Fisk, & Rogers, 2006). One other perception measure used in a study was a binary question on intended dependence (Dzindolet, Pierce, Beck, & Dawe, 2002). In the present data, the relationship between perceived reliability and system usage was weak. On the other hand, the relationship between perceptions of system use (perceived compliance, perceived reliance, perceived dependence) that directly map on to the system use variables (compliance, reliance, dependence) had a strong, albeit negative, relationship with system use. To that end, studies could potentially use perceptions of system use instead of perceived reliability as predictors and mediators of system use, including projecting future use. Perceptions of reliability could still be informative for understanding a person's perceptions of the system as well as motivation to continue using a system.

Note that the correlations between perceptions of system usage and actual system usage used eight data points per person. However, this method of analysis could be biased by individual differences. Additional correlation methods should be explored to better understand these relationships.

Implications for Practice

Knowledge of the differential effects of introduction formats can inform different introductions based on a given system's needs, including for specific tasks, domains, and goals. Overall, the results could inform design, instructions, and training programs for automation. The results could be applied in both operational environments, and could also be implemented in research studies involving automation.

Perceptions of reliability are differentially influenced by the two formats of introduction. Explicit statements have a robust impact on perceptions of system reliability, with perceptions for these groups aligning in an ordinal manner with the level of introduced expectation. On the other hand, the levels of introduced expectation for initial exposure groups did not align in an ordinal relationship with their perceived reliability; the higher-than group for this introduction format resulted in lower perceptions of reliability. For this reason, if a designer wants users to perceive a system consistent with the system's actual capabilities, they should explicitly state these capabilities. However, designers should be cautious not to allow users to initially use a demo system or similar system that has a higher reliability than the system they will be using over time.

Automation use was also differentially influenced by the two introduction formats. Initially, compliance was not different between groups, and reliance was not different between groups. Compliance and reliance within each group were also not different, except for the initial exposure lower-than group, which initially complied more than relied. Therefore, if a designer needs a user to initially comply more than rely on a system (e.g., automation misses are initially more costly than automation false alarms), the introduction should be an initial exposure to a system with lower reliability than the actual system. Overall, explicit statement groups comply with systems more than initial exposure groups. Also, patterns of reliance on the system for the different levels of expectation in initial exposure groups change over time, whereas the patterns are more robust for the different levels of expectation for explicit statement groups. To this end, if it is important for users to have higher compliance with a system over time (e.g., low cost of an initial automation false alarm), then the designer should ensure that the system is introduced with an explicit statement, regardless of the match of the level of introduced reliability to the actual system reliability.

For example, imagine a system where implementation does not occur in a strictly controlled environment. The goal of the system is for users to have high compliance, and is important for users to continue to perceive the system as having a high reliability. In this instance, using an explicit statement of a higher reliability than the actual system reliability could achieve the system's goals.

Future Directions

The next step to continue researching this space is to repeat this study with a larger sample size. Having more participants would both increase power and potentially lead to identification of individual difference trends in automation usage patterns. More stable data at the block level of time would allow for higher granularity in analyzing the data—in particular, the effects of expectations over time. Other predictors of how individual differences could impact automation use could also be used.

Additionally, as mentioned in the Perceived Reliability section of the discussion, the (lack of) impact from the initial exposure expectation format on perceptions should be explored further. To do this, participants first should report perceptions before interacting with the automation to attempt to understand their general mental model about reliability of automation. Additionally, the length of initial exposure should be manipulated to determine what length of exposure is needed to find differences between expectation levels.

Different levels of reliability could be manipulated, including varying the distances between the expected reliabilities versus the actual system reliabilities. Also, people interpret number formats (e.g., percentages, statements) differently (Fausset & Rogers, 2012), so changing how the reliability level is stated could have an impact on expectations. For both expectation format groups, the level of manipulated reliability could be systematically altered to find inflection points of where system disuse occurs.

Additionally, future automation introductions could be framed in different domains (e.g., medical) and tasks (e.g., surgery) with varying levels of immediacy (e.g., urgent vs. non-urgent) and severity of errors (e.g., life-critical vs. routine), to investigate the effect of introduction factors on the human-automation relationship. These findings could inform the model of human-automation interaction used by researchers.

APPENDIX A

<u>EXPLICIT STATEMENT DESCRIPTIONS</u>

Higher Expectancy Condition

An Automated Warehouse Management System is a system that scans the inside of truck boxes, calculates the amount of space available in the truck, loads shipments onto the truck, determines if the truck is full, and when the truck is full notifies the Supervising Warehouse Manager to dispatch the truck.

SRT-2 Automated Warehouse Management System

We are working with a company on issues of automation, as well as being funded by NIH for this work. Let me tell you a little about the system you will be helping us test. The company first became involved in sensory technologies in 1975 with the sole mission of creating advanced scanning and decision making systems for warehouse loading and shipping applications. In 1985, the company released an Automated Warehouse Management System called the SRT and in 1997, released a Smart Automated Warehouse Management System, the SRT-1. The company's latest groundbreaking system, the SRT-2, utilizes advanced decision algorithms and sensing technologies that have the ability to adjust to differing warehouse and loading conditions. Testing of the SRT-2 indicates that it is the industry standard for accuracy, reliability, and robustness and is still considered the leader in Automated Warehouse Management System systems.

Two types of errors can potentially be committed: a false alarm or a miss. A false alarm is when the system indicates that a truck is full when in fact it is not full. For example, like when smoke alarm sounds when there is no fire. A miss is when the system fails to indicate that the truck is full when in fact it IS full. For example, when there is a fire but the smoke alarm does not sound. We need your data as the baseline for other system software we are evaluating.

Because this is a well proven Automated Warehouse Management System, it is expected that the SRT-2 will perform at a high level with some performance errors. The SRT-2 has a **system reliability of <u>90%</u>**. The software running your system is the basis for the SRT-2. We need your data as the baseline for other system software we are evaluating.

Lower Expectancy Condition

An Automated Warehouse Management System is a system that scans the inside of truck boxes, calculates the amount of space available in the truck, loads shipments onto the truck, determines if the truck is full, and when the truck is full notifies the Supervising Warehouse Manager to dispatch the truck.

SRT-2 Automated Warehouse Management System

We are working with a company on issues of automation, as well as being funded by NIH for this work. Let me tell you a little about the system you will be helping us test. The company first became involved in sensory technologies in 1975 with the sole mission of creating advanced scanning and decision making systems for warehouse loading and shipping applications. In 1985, the company released an Automated Warehouse Management System called the SRT and in 1997, released a Smart Automated Warehouse Management System, the SRT-1. The company's latest groundbreaking system, the SRT-2, utilizes advanced decision algorithms and sensing technologies that have the ability to adjust to differing warehouse and loading conditions. Testing of the SRT-2 indicates that it is the industry standard for accuracy, reliability, and robustness and is still considered the leader in Automated Warehouse Management System systems.

Two types of errors can potentially be committed: a false alarm or a miss. A false alarm is when the system indicates that a truck is full when in fact it is not full. For example, like when smoke alarm sounds when there is no fire. A miss is when the system fails to indicate that the truck is full when in fact it IS full. For example, when there is a fire but the smoke alarm does not sound. We need your data as the baseline for other system software we are evaluating.

Because this is a first prototype Automated Warehouse Management System, it is expected that the SRT-2 will perform at a low level with some performance errors. The SRT-2 has a **system reliability of 60%**.

Same-As Expectancy Condition

An Automated Warehouse Management System is a system that scans the inside of truck boxes, calculates the amount of space available in the truck, loads shipments onto the truck, determines if the truck is full, and when the truck is full notifies the Supervising Warehouse Manager to dispatch the truck.

SRT-2 Automated Warehouse Management System

We are working with a company on issues of automation, as well as being funded by NIH for this work. Let me tell you a little about the system you will be helping us test. The company first became involved in sensory technologies in 1975 with the sole mission of creating advanced scanning and decision making systems for warehouse loading and shipping applications. In 1985, the company released an Automated Warehouse Management System called the SRT and in 1997, released a Smart Automated Warehouse Management System, the SRT-1. The company's latest groundbreaking system, the SRT-2, utilizes advanced decision algorithms and sensing technologies that have the ability to adjust to differing warehouse and loading conditions. Testing of the SRT-2 indicates that it is the industry standard for accuracy, reliability, and robustness and is still considered the leader in Automated Warehouse Management System systems.

Two types of errors can potentially be committed: a false alarm or a miss. A false alarm is when the system indicates that a truck is full when in fact it is not full. For example, like when smoke alarm sounds when there is no fire. A miss is when the system fails to indicate that the truck is full when in fact it IS full. For example, when there is a fire but the smoke alarm does not sound. We need your data as the baseline for other system software we are evaluating.

The Automated Warehouse Shipping Manage system that you will be interacting with today is very reliable but may make performance errors. The SRT-2 has a **system reliability of <u>75%</u>**.

APPENDIX B

AUTOMATION ATTITUDES QUESTIONNAIRE

Please mark the appropriate response

1. I feel comfortable using automated devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

2. Automation will never replace the need for working human beings.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

3. Learning to use automated devices is a worthwhile and necessary subject.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

4. Reading or hearing about automated devices would be (is) boring.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

5. Automated devices are making the jobs done by humans less important.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

6. Automated devices make me nervous.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

7. I don't care to know more about automation.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

8. Automated devices would be (are) fun to use.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

9. I don't feel confident about my ability to use automated devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

10. People are smarter than automated devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

11. Automated devices are too fast.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

12. Automation is confusing.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

13. Given a little time and training, I know I could learn to use most automated devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

APPENDIX C

AUTOMATION EXPERIENCE QUESTIONNAIRE

The purpose of this questionnaire is to assess your familiarity and experience with automated devices. Please answer all questions.

As cited in Moray, Inagaki, and Itoh (2000), automation is defined as “any sensing, detection, information-processing, decision making or control action that could be performed by humans but is actually performed by a machine” (pp. 44). A couple of everyday examples are the gas gauge in a car or the autopilot system in an airplane. Keep the above definition in mind when answering the following questions.

1. Please circle the pieces of automation (or automated devices) you use <u>on a regular basis</u>.

Computer	Personal Data Assistant (PDA)	Cruise control
Blood Glucose Meter	Washing and/or Drier	Remote Control
Game System (e.g., X-Box)	Thermostat	Copy Machine
Hearing aids	Alarm clock	Scale
Car seat adjustment	Hair drier	Vacuum
Voicemail	VCR	DVD/CD Player
Answering machine	Iron	Cell Phone
ATM machine	Other(s): ______________________	

2. Please circle the pieces of automation you use <u>on an infrequent basis.</u>

Computer	Personal Data Assistant (PDA)	Cruise control
Blood Glucose Meter	Washing and/or Drier	Remote Control
Game System (e.g., X-Box)	Thermostat	Copy Machine
Hearing aids	Alarm clock	Scale
Car seat adjustment	Hair drier	Vacuum
Voicemail	VCR	DVD/CD Player
Answering machine	Iron	Cell Phone
ATM machine	Other(s): ______________________	

3. Please list two automated devices/systems you find the MOST DIFFICULT to use.

4. For what reason or reasons do you have difficulty using these devices/systems?

5. Please list two automated devices/systems you find the EASIEST to use.

6. For what reason or reasons do you find it easy to use these devices/systems?

Using the scale below, circle the answer in question 7 that best represents your opinion

7. Overall, how much do you trust automated devices/systems?

1	2	3	4	5
Not at all				Completely

APPENDIX D

Automation-Induced Complacency Potential Questionnaire

1. Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

2. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because computerized surgery is more reliable and safer than manual surgery.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

3. People save time by using automatic teller machines (ATMs) rather than a bank teller in making transactions.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

4. I do not trust automated devices such as ATMs and computerized airline reservation systems.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

5. People who work frequently with automated devices have lower job satisfaction because they feel less involved in their job than those who work manually.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

6. I feel safer depositing my money at an ATM than with a human teller.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

7. I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my VCR rather than manual taping.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

8. People whose jobs require them to work with automated systems are lonelier than people who do not work with such devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

9. Automated systems used in modern aircraft, such as the automatic landing system, have made air journey safer.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

10. ATMs provide safeguard against the inappropriate use of an individual's bank account by dishonest people.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

11. Automated devices used in aviation and banking have made work easier for both employees and customers.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

12. I often use automated devices.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

13. People who work with automated devices have greater job satisfaction because they feel more involved than those who work manually.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

14. Automated devices in medicine save time and money in the diagnosis and treatment of disease.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

15. Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed-trap in case the automatic control is not working properly.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

16. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

17. I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

18. Work has become more difficult with the increase of automation in aviation and banking.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

19. I do not like to use ATMs because I feel that they are sometimes unreliable.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

20. I think that automated devices used in medicine, such as CAT-scans and ultrasound, provide very reliable medical diagnosis.

1	2	3	4	5
Strongly Disagree	Agree	Neither agree nor disagree	Disagree	Strongly Agree

APPENDIX E

DEMOGRAPHICS AND HEALTH QUESTIONNAIRE

Please answer the following questions. All of your answers will be treated confidentially. Any published document regarding these answers will not identify individuals with their answers. **If there is a question you do not wish to answer, please just leave it blank and go on to the next question.** Thank you in advance for your help.

Gender: Male ☐$_1$ **Female** ☐$_2$ **Age: _______**

1. Do you consider yourself Hispanic or Latino?

☐$_1$ Yes

1 a. If "Yes", would you describe yourself:

☐$_1$ Cuban
☐$_2$ Mexican
☐$_3$ Puerto Rican
☐$_4$ Other (please specify) ________________

☐$_2$ No

2. How would you describe your primary racial group?

☐$_1$ No Primary Group
☐$_2$ White Caucasian
☐$_3$ Black/African American
☐$_4$ Asian
☐$_5$ American Indian/Alaska Native
☐$_6$ Native Hawaiian/Pacific Islander
☐$_7$ Multi-racial
☐$_8$ Other (please specify) ______________________

3. Are you fluent in English?

☐$_1$ Yes
☐$_2$ No

4. Is English your native language?

☐$_1$ Yes
☐$_2$ No

4a. If "No", What is your primary language? _____________________

Medication Information Form

Please list the medical products that you are currently taking. Include medicinal herbs, vitamins, aspirin, etc., as well as prescription medications (copy names from label if possible).

Below is an example of how to fill out the form. If you take Ibuprofen for Arthritis two times a day, you would fill the form out as shown in the example below. There is space for up to eight different medications. If you take more than eight medications regularly, please list the rest on the back of the last page.

Name of Medication	Reason for taking medication	How often do you take this medication? (Please select one)
Example: **Ibuprofen**	**Arthritis**	☒ Daily __2__ times/day ☐ Weekly ______ times/week ☐ Monthly ______ times/month ☐ As Needed

Please turn the page to list your medications

Name of Medication	Reason for taking medication	How often do you take this medication? (Please select one)
1.		☐ Daily ______ times/day ☐ Weekly ______ times/week ☐ Monthly ______ times/month ☐ As Needed
2.		☐ Daily ______ times/day ☐ Weekly ______ times/week ☐ Monthly ______ times/month ☐ As Needed
3.		☐ Daily ______ times/day ☐ Weekly ______ times/week ☐ Monthly ______ times/month ☐ As Needed
4.		☐ Daily ______ times/day ☐ Weekly ______ times/week ☐ Monthly ______ times/month ☐ As Needed

APPENDIX F

EXPECTANCY QUESTIONNAIRE

Please circle the number that corresponds to how well you expect the Automated System to perform on the upcoming task.

1	2	3	4	5
Not well at all				**Perfectly**

1. Please indicate how often you believe the Automated System will provide correct information (using a %. 0% being never and 100% being always). (Example: I think the Automated System will be correct ##% of the time)

   _____%

1. Please indicate how much you plan to rely on the Automated System (using a %).
   (Example: I plan to rely on the Automated System ## % of the time)

   _____%

2. Please circle the number that corresponds to the likelihood of the Automated System committing an error.

1	2	3	4	5
Not at all likely				**Extremely likely**

3. Please indicate how you perceive the relationship between automated systems and human users.

1	2	3	4	5
Automation works for the human		**Collaborative Team**		**Human works for the automation**

APPENDIX G

INTERIM QUESTIONNAIRE

1) Please indicate, using a percentage, how often you thought the automated system alerted you <u>at the correct time</u> to dispatch a full truck (0-100%)?

   _____________%

2) How much did you <u>trust</u> the automated system to correctly alert you when a truck was full and ready to be dispatched?

1	**2**	**3**	**4**	**5**
Not at all		**Neutral**		**Completely**

3) What percentage of the time did you view the truck when the automated system gave you an alert to dispatch the truck (0-100%)?

   _____________%

4) What percentage of the time did you view the truck when there was no alert present from the automated system (0-100%)?

   _____________%

APPENDIX H

AUTOMATION ERRORS IN THE MANIPULATION BLOCK OF 12 TRUCKS FOR THE INITIAL EXPOSURE GROUPS

	System Reliability for Manipulation Block (%)	Group	# Misses out of 12 Trucks	# False Alarms out of 12 Trucks
Higher	91.7	Group A	1	0
		Group B	0	1
Same-As	75.0	Group A	2	1
		Group B	1	2
Lower	58.3	Group A	3	2
		Group B	2	3

APPENDIX I

AUTOMATION ERRORS IN THE EXPERIMENTAL BLOCKS OF 20 TRUCKS FOR ALL CONDITIONS

	System Reliability for Experiment Block (%)	# Misses out of 20 Trucks	# False Alarms out of 20 Trucks
Block 1	75	3	2
Block 2	75	2	3
Block 3	75	2	3
Block 4	75	3	2
Block 5	75	3	2
Block 6	75	2	3
Block 7	75	2	3
Block 8	75	3	2

APPENDIX J

AUTOMATION ERRORS TIMING AND TYPES FOR ALL MANIPULATIONS, BLOCKS, AND CONDITIONS

Explicit Statement (all)		N/A	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss		11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm		16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-H (A)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss	7	11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm		16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-H (B)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss		11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm	6	16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-SA (A)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss	7,12	11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm	6	16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-SA (B)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss	7	11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm	6,9	16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-L (A)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss	7,12,2	11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm	6,9	16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

IE-L (B)		Manipulation	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
	Miss	7,12	11,2,20	11,9	6,19	10,6,2	9,15,20	10,17	1,10	7,8,13
	False Alarm	6,9,2	16,13	5,8,15	14,16,3	16,20	7,12	18,16,2	14,16,8	19,12

REFERENCES

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688-699.

Barg-Walkow, L. H., Walsh, D. R., & Rogers, W. A. (2012). Understanding use errors for medical devices: Analysis of the MAUDE database. *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (pp. 872-876). Santa Monica, CA: Human Factors and Ergonomics Society.

Bliss, J., Dunn, M., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills, 80*(3c), 1231-1242.

Bibby, P. A., & Payne, S. J. (1993). Internalization and the use specificity of device knowledge. *Human-Computer Interaction, 8*, 25-56.

Bowers, C. A., Oser, R. L., Salas, E., & Cannon-Bowers, J. A. (1996). Team performance in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum.

Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology, 35*, 38-48.

Chappell, S. L. (1997). The effects of experience and automation on failure detection. *Proceedings of the 9th International Symposium on Aviation Psychology* (pp. 773-779). Columbus, OH: The Ohio State University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *Proceedings of the 1998 Command and Control Research and Technology Symposium* (pp. 1–37). Washington, DC: Department of Defense C4ISR Cooperative Research Program.

Czaja, S. J., Charness, N., Dijkstra, K., Fisk, A. D., Rogers, W. A., & Sharit, J. (2006). *Center for Research and Education on Aging and Technology Enhancement: Demographic and Background Questionnaire* (CREATE-TR-2006-02). University of Miami: CREATE.

Dixon, S. & Wickens, C. D. (2004). *Reliability in automated aids for unmanned aerial vehicle flight control: Evaluating a model of automation dependence in high workload* (AHFD-04-05/MAAD-04-1). Savoy, Illinois: University of Illinois at Urbana-Champaign, Aviation Human Factors Division, Institute of Aviation.

Dixon, S. R. & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48,* 474-486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007) On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49,* 564-572.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*, 697-718.

Dzindolet, M. T., Pierce, L. G., & Beck, H. P. (2009). An Examination of the Social, Cognitive, and Motivational Factors that Affect Automation Reliance. *Human Factors Issues in Combat Identification*, 277-297.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79-94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, *13*(3), 147-164.

Eiriksdottir, E. (2011). *The role of principles in instructions for procedural tasks: timing of use, method of study, and procedural instruction specificity* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors, 50*, 853-863.

Fausset, C. B., & Rogers, W. A. (2012). Younger and older adults' comprehension of health risk probabilities: Understanding the relationship between format and numeracy. *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (pp. 120-124). Santa Monica, CA: Human Factors and Ergonomics Society.

Jamieson, D. W., Lydon, J. E., Stewart, G, & Zanna, M. P. (1987). Pygmalion revisited: New evidence for student expectancy effects in the classroom. *Journal of Educational Psychology, 79*, 461-466.

Johnson, J. D. (2004). *Type of automation failure: The effects on trust and reliance in automation*. (Unpublished master's thesis). Georgia Institute of Technology, Atlanta.

Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*, 191-219.

Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science, 47*(5), 693-706.

Lee, J. D. (2006). Human factors and ergonomics in automation design. *Handbook of Human Factors and Ergonomics, Third Edition*, 1570-1596.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153-184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50-80.

Madhavan, P., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators' utilization of diagnostic advice. *Proceedings of the*

Human Factors and Ergonomics Society 49th Annual Meeting (pp. 487-491). Santa Monica, CA: Human Factors and Ergonomics Society.

Madhavan, P. & Wiegmann, D. A. (2007). Effects on information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors, 49*, 773-785.

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors, 45*(2), 281-295.

Masalonis, A. J., & Parasuraman, R. (1999, September). Trust as a construct for evaluation of automation aids: Past and future theory and research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 184-187). Santa Monica, CA: Human Factors and Ergonomics Society.

Mayer, A. K. (2006). *The manipulation of user expectancies: Effects on reliance, compliance, and trust using an automated system*. (Unpublished master's thesis). Georgia Institute of Technology, Atlanta, Georgia.

Mayer, A. K., Fisk, A. D., & Rogers, W. A. (2008). Age differences, expectancy, and resultant reliance and compliance behavior using an automated system. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.

Mayer, A. K., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2006). Don't let me down: The role of operator expectations on human-automation interaction. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 2345-2349). Santa Monica, CA: Human Factors and Ergonomics Society.

McBride, S. E. (2010). *The effect of workload and age on compliance with and reliance on an automated system*. Unpublished master's thesis, Georgia Institute of Technology, Atlanta, Georgia.

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the Effect of Workload on Automation Use for Younger and Older Adults. *Human Factors, 53*(6), 672-686.

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2013). Understanding Human Management of Automation Errors. *Theoretical Issues in Ergonomics Science*. Advance online publication. doi: 10.1080/1463922X.2013.817625

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44-58.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201-220). Mahwah, NJ: Lawrence Erlbaum.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, 1905-1922.

Norman, D. A. (1983). Some observations on mental models. In A. L. Stevens & D. Gentner (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1987). Some observations on mental models. In R. M. Baecker & W. A. S. Buxton (Eds.), *Readings in Human-Computer Interaction: A Multidisciplinary Approach* (pp. 241-244). Los Altos, CA: Morgan Kaufmann.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans*, *30*, 286-297.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, *50*(3), 511-520.

Pop, V. L. (2012). *Individual differences in the calibration of trust in automation*. (Unpublished master's thesis). Georgia Institute of Technology, Atlanta.

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 19-35). Mahwah, NJ: Lawrence Erlbaum.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart, and Winston.

Russell, B. (1910) Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society* 11: 108-128. Reprinted in Russell, B. (1963) *Mysticism and Logic* (152-167). London: Allen and Unwin.

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta.

Sanchez, J. (2009,). Conceptual Model of Human-Automation Interaction. *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (pp. 1403-1407). Santa Monica, CA: Human Factors and Ergonomics Society.

Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 586-589). Santa Monica, CA: Human Factors and Ergonomics Society.

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science, 2*(5), 341-350.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993a). Automation-induced "complacency": Development of a complacency-potential scale. *International Journal of Aviation Psychology, 3*, 111-122.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991-1006.

Stephan, W. G. (1985). Intergroup relations. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 2, pp. 599-658). New York, NY: Random House.

Wasserman, E.A. (l990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science, 1*(5), 298-302.

Wickens, C. D., & Carswell, C. M. (2006). Information processing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics: Third Edition* (pp. 1570-1596). Hoboken, NJ: John Wiley & Sons, Inc.

Wickens, C., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors, 49*(5), 839-850.

Wickens, C. D., & Dixon, S. (2002) *Workload demands of remotely piloted vehicle supervision and control: (I) Single vehicle performance* (AHFD-02-10/MAD-02-1). Savoy, Illinois: University of Illinois at Urbana-Champaign, Aviation Human Factors Division, Institute of Aviation

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201-212.

Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors, 44,* 171-188.

Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention* (AHFD-02-14/MAAD-02-2). Savoy, Illinois: University of Illinois at Urbana-Champaign, Aviation Human Factors Division, Institute of Aviation.